ScuDo

Scuola di Dottorato ∽ Doctoral School

WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation

Doctoral Program in Electrical, Electronics and Communications Engineering

(30^{th}cycle)

Power Management Circuits for Front-End ASICs

Employed in High Energy Physics Applications

By

Junying Chai

******

Supervisor(s):

Prof. A. Rivetti, Supervisor
Prof. S. Marcello, Co-Supervisor
Dr. M. Da Rocha Rolo, Co-Supervisor

Doctoral Examination Committee:

Prof. Valerio.Re, Referee, University of Bergamo, Italy
Prof. Michela.Chiosso, Referee, University of Turin, Italy
Prof. E.F, University of...
Prof. G.H, University of...
Prof. I.J, University of...

Politecnico di Torino
2018

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Junying Chai
2018

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

I would like to dedicate this thesis to my loving parents Dongwen Chai and Fengzhi Zhao, my loving wife Yanying Zheng, my loving son Haoze Chai and the imminent child.

Acknowledgements

I would like to acknowledge my tutors, my colleagues and my friends.

My gratitude is sincere.

Marco Maggiora helped me driving my work at the INFN Torino in Italy.

Angelo Rivetti, one of my supervisor, guided my doctoral project and research direction.

Simonetta Marcello, as one of my supervisor gave me a lot of help in my study in Politecnico and my life in Italy.

Mannel Rolo, as one of my supervisor appreciated me to improve my skills in developing analogue electronics and gave me good advices.

I also thank Ennio, Gianni, Lino, Huashen, Chongyang who gave me many help in my study and research.

Abstract

The instrumentation of radiation detectors for high energy physics calls for the development of very low-noise application-specific integrated-circuits and demanding system-level design strategies, with a particular focus on the minimisation of interference noise from power management circuitry. On the other hand, the aggressive pixelisation of sensors and associated front-end electronics, and the high radiation exposure at the innermost tracking and vertex detectors, requires radiation-aware design and radiation-tolerant deep sub-micron CMOS technologies.

This thesis explores circuit design techniques towards radiation tolerant power management integrated circuits, targeting applications on particle detectors and monitoring of accelerator-based experiments, aerospace and nuclear applications. It addresses advantages and caveats of commonly used radiation-hard layout techniques, which often employ Enclosed Layout or H-shaped transistors, in respect to the use of linear transistors.

Radiation tolerant designs for bandgap circuits are discussed, and two different topologies were explored. A low quiescent current bandgap for sub-1 *V* CMOS circuits is proposed, where the use of diode-connected MOSFETs in weak-inversion is explored in order to increase its radiation tolerance.

An any-load stable LDO architecture is proposed, and three versions of the design using different layout techniques were implemented and characterised.

In addition, a switched DC-DC Buck converter is also studied. For reasons concerning testability and silicon area, the controller of the Buck converter is on-chip, while the inductance and the power transistors are left on-board.

A prototype test chip with power management IP blocks was fabricated, using a TSMC 65 *nm* CMOS technology. The chip features Linear, ELT and H-shape

LDO designs, bandgap circuits and a Buck DC-DC converter. We discuss the design, layout and test results of the prototype.

The specifications in terms of voltage range and output current capability are based on the requirements set for the integrated on-detector electronics of the new CGEM-IT tracker for the BESIII detector. The thesis discusses the fundamental aspects of the proposed on-detector electronics and provides an in-depth depiction of the front-end design for the readout ASIC.

Key words: LDO, Bandgap, DC-DC, ASIC, CMOS, Power Management, High Energy Physics.

Contents

List of Figures

List of Tables

Nomenclature

Roman Symbols

C_{ox}	the specific gate capacitance
p_{em}	emission probability of an electron from the valence band of $\mathrm{S_iO_2}$ to the traps
T	absolute temperature
X_{m}	density of the holes at the depth of m
A	relative atomic mass
c	speed of light
e	electron charge
k_B	Boltzmann constant
m_e	rest mass of electron
Mu	Molar mass constant
N_A	Avogadro number
q	an electron charge $1.6 \times 10^{-19}\ C$
Z	atomic number
X	X-rays

Greek Symbols

β	$\frac{v}{c}$, relative speed for light

ΔV_{in} threshold shift caused by the interface state

ΔV_{ox} threshold shift caused by the gate oxide

ΔN density of the trapped holes

ε_0 vacuum permittivity

γ gamma rays

ρ density of the material

φ energy difference between the trap level and the valence band

e-h electron-hole pair

Acronyms / Abbreviations

AC Alternating Current

AMS Alpha Magnetic Spectrometer

ASIC Application-Specific Integrated Circuit

ATICE A Large Ion Collider Experiment

ATLAS A Toroidal LHC Apparatus

BEPCII Beijing Electron–Positron Collider II

BeppoSAX Satellite per Astronomia X, Beppo in honer of Giuseppe Occhialini

BESIII Beijing Electron Spectrometer III

BJT Bipolar Junction Transistor

CEPC Circular Electron Positron Collider

CERN The European Organization for Nuclear Research

CGEM Cylindrical Gas Electron Multiplier

CGEM-IT Cylindrical Gas Electron Multipliers Inner Tracker

CMC Current Mode Control

CMS	Compact Muon Solenoid
CTAT	Complementary to Absolute Temperature
DAMPT	Department of Applied Mathematics and Theoretical Physics
DC	Direct Current
DC-DC	Direct Current to Direct Current
DCR	Direct Current Resistance
DTMOST	Dynamic Threshold MOSFET
EMC	Electro-Magnetic Calorimeter
ESR	Equivalent Series parameter Resistor
EXOSAT	European X-ray Observatory Satellite
FEE	Front-End Electronics
FIT	The Failure In Time
HEP	High Energy Physics
HERD	High Energy cosmic Radiation Detection facility
HXMT	Hard X-ray Modulation Telescope
IHEP	Institute of High Energy of Physics Chinese Academy of Sciences
INFN	Istituto Nazionale di Fisica Nucleare
IS	International System
KCL	Kirchhoff’s Circuit Laws
KCL	Kirchhoff’s Circuit Laws
KEK	The High Energy Accelerator Research Organization
LDO	Low Drop Out Regulator
LET	The Linear Energy Transfer

LHC	Large Hadron Collider
LHCb	LHC beauty
LOCOS	LOCal Oxidation of Silicon
MC	Monte Carlo simulation
MDC	Main Drift Chamber
MRI	Magnetic Resonance Imaging
MTBF	The Mean Time Between Failures
MUC	Muon Chamber
NATO	North Atlantic Treaty Organization
NIEL	Non-Ionizing Energy Loss
OPA	OPeration Amplifier
PET	Positron Emission Yomography
PID	Particle Identification
PMT	PhotoMultiplier Tubes
PS	Proton Synchrotron
PSRR	Power Supply Ripple Rejection
PTAT	Proportional to Absolute Temperature
SEE	Single Event Effects
SOI	Silicon On Insulator
SRAM	Static Random Access Memory
STI	Shallow Trench Isolation
TC	Temperature Coefficient
TID	Total Ionizing Dose

TIGER	Torino Integrated GEM Electronics for Readout
TMR	Triple Module Redundancy
TOF	Time Of Flight system
VMC	Voltage Mode Control

Chapter 1

The Development of the Front-End ASICs in High Energy Physics

1.1 High Energy Physics Research

The object of high energy physics is to research the high energy photons and particles. One of the purposes is to research the origin of the world. For example, what is the matter composed of, how does the mass produce and so on. The normal research method is using an experimental application to detect the high energy photons and particles directly or indirectly.

Normally, the high energy photons and particles have two kinds of sources. One source comes from the high energy accelerator with the collider, the other source comes from the space. Up to now, the biggest and the highest energy collider is the LHC (Large Hadron Collider) at CERN, as shown in Figure 1.1. There are four experiments: CMS, ALICE, ATLAS and LHCb.

The LHC is about 27 kilometers long and 100 meters under the ground. The protons are accelerated up to 1.4 GeV, and through the Proton Synchrotron (PS), they can be accelerated up to 26 GeV. In 2012, the Higgs boson particle has been detected for the first time at LHC.

Other high energy accelerators with experimental application are for instance the BEPCII with BESIII (at IHEP, Beijing, China), the SuperKEKB with BELLEII (at KEK, Tsukuba, Japan), and so on. The higher energy accelerators experiments

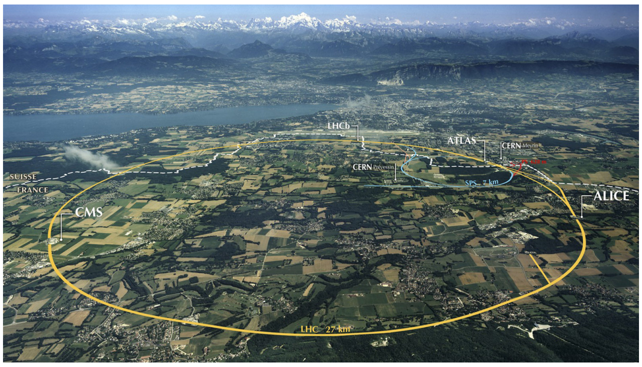


Fig. 1.1: LHC including the CMS, ALICE, LHCb and ATLAS detectors

planed in the near future include the CEPC in China and the ILC in Japan, as shown in Figure 1.2.

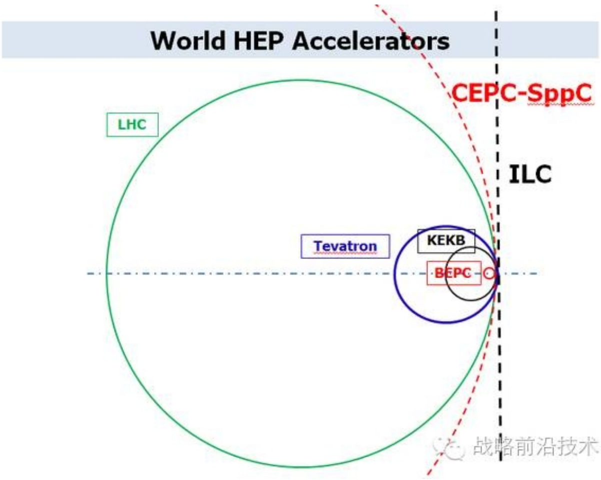


Fig. 1.2: Comparision of CEPC, ILC and the present accelerators

Besides the collider, there are also other kinds of physics applications, such as the synchrotron radiation source, the spallation neutron source and so on. High energy physics applications benefits all of us in many aspects, especially in the nuclear medicine such as the magnetic resonance imaging (MRI), the positron emission tomography (PET), the Gamma Knife, the Proton knife and so on.

High energy physics investigates also the high energy photons and particles coming from the space. The main purpose is the history of the evolution and the origin of the universe.

Because of the presence of the high energy photons and particles in the cosmic rays, this kind of research does not need accelerates and just need detectors. Research can be classified space detectors, ground detectors and underground detectors.

For space detectors, the platform includes the space station, satellite, and the ball based. Such as the AMS, HXMT, DAMPE and so on. The AMS is shown in Figure 1.3.

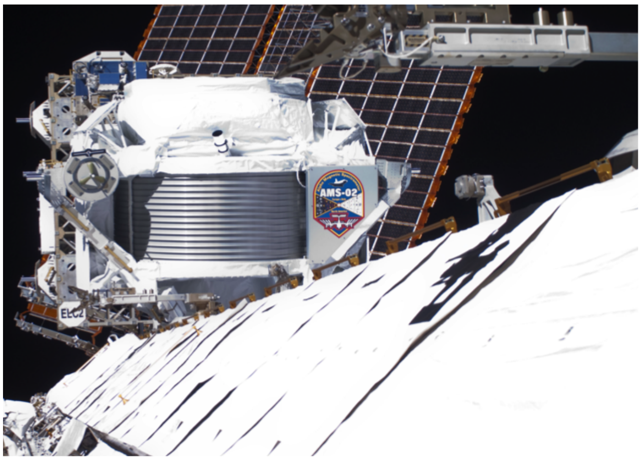


Fig. 1.3: AMS installed in the International Space Station

The newest detector facing the space high energy photons and particles is the HXMT satellite, which is made by IHEP, Beijing, China. The HXMT has been launched in June, 2017, and is shown in Figure 1.4. The HXMT can detect the low energy, the middle energy and the high energy X rays from 1 keV up to the 250 keV.

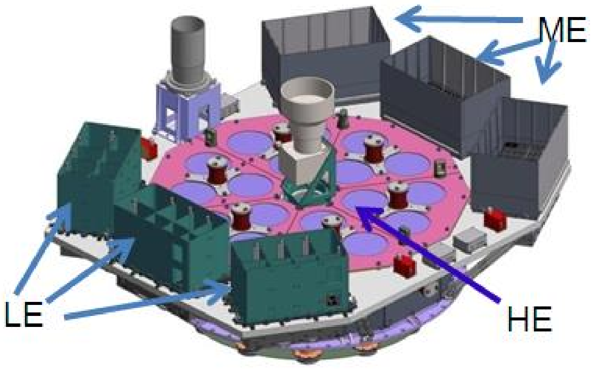


Fig. 1.4: Illustration of the payload configuration on the HXMT satellite

The ground detectors are mainly built on the high plateau area where the air is thin and dry. Such as the ARGO which is in the Tibet, China, which can detect the high energy particle indirectly.

The underground detectors are built in very deep places, the main purpose is to detect the high energy particles which can be filtered by the mountain and the ground, such as the neutrinos, the dark matter particles and so on.

The development of high energy physics requires the use of the detectors and the corresponding front-end readout electronics. Because of each is unique, the front-end electronics is not available in the market. This need to design a custom Application-Specific Integrated Circuit (ASIC) to be used in the detectors. Next section is about the front-end ASIC suitable for high energy physics detectors.

1.2 Power management status in front-end ASICs

By investigating the ASIC of high-energy physics front-end electronics, it is found that most of the ASIC chips do not integrate power management and use external power supply. Comparing with the external power supply, the on-chip power management system has three advantages:

The first advantage is that it can improve the load respond ability of the front-end electronic. Off-chip power supply will have the parasitic inductance brought by the pins and bonding wires, which is about between a few nH to a few tens of nH. When the front-end electronic current changes rapidly, not only the interference voltage is generated by the parasitic inductance, but also the reaction speed is delayed. The on-chip power supply is more suitable in such situations.

The second advantage is that the power supply area can be reduced. The on-chip power supply system can save the additional package area, so it is suitable for area sensitive applications.

The third advantage is that the common power management chips have poor radiation resistance. Even if there is the commercial shelf chip with good anti-radiation capability, it will be expensive, long delivery period or difficult to buy (later discussing further). Designing one new power management can integrate radiation-aware methods.

In order to understand the current state of power management in front-end electronics in the field of high energy physics, this thesis researched the main high energy physics experiments. The following three tables are for the ASICs being used and ready to be used in some of the world's major high energy physics experiments.

Table 1.1: Power management comparison table 1 in front-end ASICs of high energy physics[3]

Experiment	Subsystem	Name	Description	Frontier	Institution	Type	Technology	**Power**	State
ATLAS	pixel	FE-I3	pixel front end chip	Energy	LBNL	mixed	250*nm* CMOS	**Outer**	Running experiment
ATLAS	pixel	FE-I4	pixel front end chip	Energy	LBNL	mixed	130*nm* CMOS	**Inner**	**Approved experiment**
ABCD	strips	ABCD	strip front end chip	Energy	UCSC	mixed	0.8*um* DMILL	**Outer**	Running experiment
ATLAS	strips	ABCn	strip front end chip	Energy	UCSC,Penn	mixed	250*nm* CMOS	**Inner**	**Candidate for Approved experiment**
ATLAS	upgrade	ABC-130	strip front end chip	Energy	Penn,UCSC	mixed	130*nm* CMOS	**Outer**	Candidate for Approved experiment
ATLAS	upgrade	HCC-130	strip front end chip	Energy	Penn,UCSC	mixed	130*nm* CMOS	**Outer**	Candidate for Approved experiment
ATLAS	TRT	ASDBLR	straw front end	Energy	Penn	analog	0.8*um* DMILL	**Outer**	Running experiment
ATLAS	Muon Samll Wheel	VMM	front end	Energy	BNL	mixed	130*nm* CMOS	**Outer**	Approved experiment
ATLAS	Muon CSC	ASM1	preamp	Energy	-	analog	0.5*um* CMOS	**Outer**	Running experiment

Table 1.3: Power management comparison table 2 in front-end ASICs of high energy physics[3]

Experiment	Sub-system	Name	Description	Frontier	Institution	Type	Technology	**Power**	State
PHENIX	strips	SVX4	strip front end chip	Nuclear Pysics	LBNL,FNAL	mixed	250*nm* CMOS	**Outer**	Running experiment
PHENIX	strips	FPHX	strip front end chip	Nuclear Pysics	FNAL	mixed	250*nm* CMOS	**Outer**	Running experiment
CLAS12	strips	FSSR2	strip front end chip	Nuclear Pysics	FNAL	mixed	250*nm* CMOS	**Outer**	Approved experiment
CMS& BelleII	strips	FSSR2	strip front end chip	Energy/intensity		mixed	250*nm* CMOS	**Outer**	Approved experiment
CMS	pixel	PSI46	pixel front end chip	Energy	-	mixed	250*nm* CMOS	**Outer**	Running experiment
CMS	ECAL	FENIX	front end, trig&DAQ	Energy	-	analog	250*nm* CMOS	**Outer**	Running experiment
FERMI	Calorimeter	GCFE	calorimeter front end	Cosmic	SLAC	analog	500*nm* CMOS	**Outer**	Running experiment
FERMI	Tracker	GTFE	strip front end chip	Cosmic	UCSC,SLAC	mixed	500*nm* CMOS	**Outer**	Running experiment
FERMI	Anti Coincidence	GAFE	PMT front end	Cosmic	SLAC	analog	500*nm* CMOS	**Outer**	Running experiment
nEXO	TPC	nEXO-FE	front end	Intensity	SLAC	analog	180*nm* CMOS	**Outer**	Candidate for proposed experiment

Table 1.4: Power management comparison table 3 in front-end ASICs of high energy physics[3]

Experiment	Sub-system	Name	Description	Frontier	Institution	Type	Technology	**Power**	State
LHC	pixel	CHIPIX65	pixel front end chip	Energy	INFN	mixed	65*nm* CMOS	**Inner**	**Approved experiment**
KLOE-2	strips	GASTONE64	strip front end chip	Energy	INFN	mixed	130*nm* CMOS	**Outer**	Running experiment
Belle II	Muon System	TARGET6B	scin strip/MPPC readout	intensity	Hawaii	mixed	250*nm* CMOS	**Outer**	Approved experiment
CTA	Camera trig/readout	TARGET5/7	scin strip/MPPC/MA PMT readout	cosmic	Hawaii	mixed	250*nm* CMOS	**Outer**	Candidate for proposed experiment

From these tables above, it can be seen that the main front-end electronics ASIC chips running in high-energy physics experiments are off-chip power management systems (red fonts). Among the next-generation experiments, there are three on-chip power management systems for front-end electronics ASICs, integrating DC/DC and LDO circuits (blue fonts) on-chip. The table reflects that the integrating power management system on the front-end chips is a relatively new direction and is a field worthy of research and exploration.

1.2.1 Radiation-aware Demands in High Energy Physics Applications

Because of the main purpose of high energy physics application, which is detecting the photons and particles, the radiation effect is inevitable. Different surroundings have different radiation-aware demands. So the front-end design must consider the special radiation environment.

Before the design of the electronics, the demand of the radiation must be clear. Some radiation effect depends on the continuous working time, some radiation effect are random but depends on the LET. For example, the applications used in space must consider the radiation from the sun, the cosmic rays, and so on, different orbits (height and inclination) have different radiation-aware demand. The LHC at CERN demands the anti-radiation effect up to 1 $Grad$(Si) and 10^{16} $Neutrons/cm^2$ in 10 years.

From the applications environment, the kinds of high energy particles and photons should be clear, and the quantity and density state can also be gotten from the simulation or from real measurement data. According to the real radiation situation, what kinds of radiation-aware components and what kinds of ASIC technology can be selected.

Another important situation should be taken into account. The United States has a national law, which is the "Export Administration Regulations" of the Bureau of Industry and Security. This national law forbids the American components, which have the above 100 $krad$(Si) anti-radiation ability, to export to the non-NATO organization including P.R.C. But most of the radiation-aware components comes from the United States. So when buying these components from the United States and using it the projects in P.R.C, there will be some problems even through the third part country.

1.3 Context and Motivation

The radiation-aware power management circuit for the front-end ASIC is motivated by the need of the CGEM-IT detector which is shown in Figure 1.5. It is used in the BESIII project at IHEP, Beijing, China. The CGEM-IT multichannel front-end ASIC is named TIGER and it is made using the UMC 110 *nm* technology.

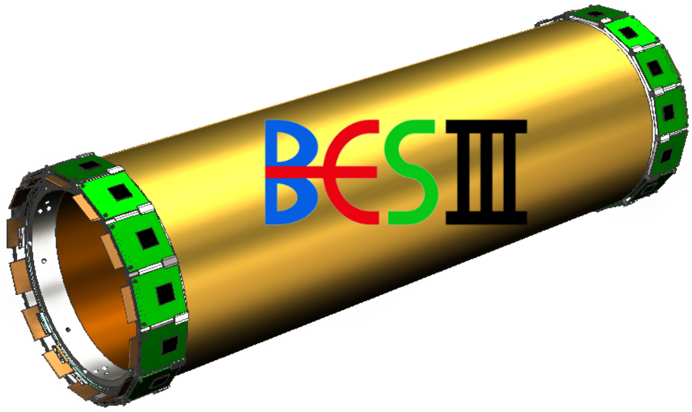


Fig. 1.5: Cylindrical GEM detector Inner Tracker (CGEM-IT)

When designing this CMOS front-end circuit, there was not enough space for the LDO on-chip and just integrated the voltage and current bias circuit. So the radiation-aware components are needed for the power supply of the TIGER, as shown in Figure 1.6. But because of the USA law mentioned in the last section, the high-quality anti-radiation power supply components cannot be imported to P.R.C., even through the third country. So it is necessary to design the radiation-aware power supply ASICs.

At the same time, the voltage and current bias circuit in the first version TIGER have not been designed properly, affecting the baseline holder. The result is the baseline voltage uncontrolled shift. So the power management including the power

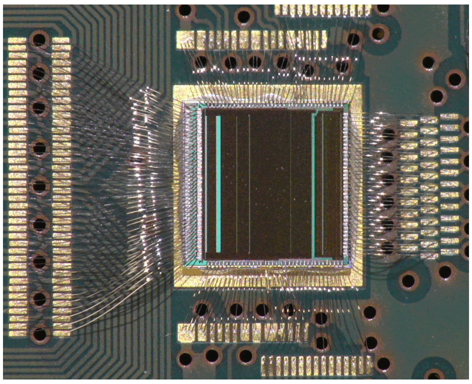


Fig. 1.6: TIGER chip bonded to the board

supply and bias circuit should be investigated carefully. The result of the research can be used in the TIGER likely front-end circuits in high energy physics applications.

The CHIPIX65 project is another motivation for this power management design. This project uses the TSMC 65 *nm* technology, and its purpose is to exploit the 65 *nm* CMOS advanced technology used in the new generation pixel detectors. This innovative 65 *nm* CMOS chip will be used in the experiment with extreme particle rates and radiation at future High Energy Physics colliders.

This CHIPIX65 project gives the opportunity to tape out for verifying the power management CMOS design on-chip, and the new power management chip is the CHIPIX65-LDO-BUCK. If this kind of power management is successful, it not only can be used in the CHIPIX65 project but also can be used to other similar projects.

The TSMC 65 *nm* technology has another superiority about the radiation-aware design. That is the radiation transistor model. This model is made by Mohsine Menouni and his colleagues [4] at CERN, and though this radiation transistor model the 200 $Mrad$(Si) and 500 $Mrad$(Si) Total Ionizing Dose (TID) radiation effect can be simulated.

1.4 Thesis Outline and Original Contributions

the main content of the thesis is designing one radiation-aware power management ASIC named CHIPIX65-LDO-BUCK, which can supplies the TIGER front-end chip used in the BESIII experiment. And this thesis deals with the design by TSMC 65 *nm* technology, which includes three kinds of LDOs, two kinds of bandgaps, one kind of switch DC-DC and one current source. The three kinds of LDOs use three kinds of NMOS transistors: the standard transistor, the H shape transistor and the ELT transistor. The two kinds of bandgaps use two different circuit structures, which can used in the advanced technology. The DC/DC type is buck, and it works together with three LDOs, two bandgaps, to get one good power management for the front-end circuit.

This thesis consists of seven chapters and one appendix.

The Chapter 1 introduces the development of high energy physics applications and the main stream front-end ASICs are introduced. Then the power management system, used in such front-end ASICs based on the features of high energy physics, is introduced. At last, the context and motivation for this thesis are introduced.

The Chapter 2 mainly focuses on the radiation effect on CMOS technology such as the TID, SEE. Based on this analysis, some radiation-aware methods are introduced. This chapter offers a new kind of shape transistor, the H shape transistor, which has a good radiation-aware ability. This transistor has some advantages such as the smaller size, and the better symmetry. Although the radiation-aware ability of H shape transistor is less than the Enclosed Layout Transistor (ELT), it is enough for the most radiation applications.

The Chapter 3 introduces the bias and the power management issues in the multichannel front-end. As the nuclear detector channels developing, the multichannel front-end becomes the mainstream. This needs the special bias circuit including the voltage and current bias. After that, the power supply to the front-end is described, which provides different kinds of power supplies. At the end, the auxiliary issues in the front-end are studied.

The Chapter 4 goes through the building blocks of the CHIPIX-LDO-BUCK: the radiation-aware LDO and bandgap. Differently from the classical LDO, this chapter uses the new architecture to design the LDO, which is suitable for the analogue digital mixed circuit. Then the bandgaps are described, and a new type of bandgap is introduced, which does not use the OPA, so the bandgap saves a lot of power consumption.

The Chapter 5 talk about the BUCK block of the power management. The BUCK is the most used switch DC-DC conversion in the front-end, and this chapter designs the Pulse Width Modulation (PWM) generator, the comparator with the compensation circuit. These circuits can work together with the off-chip inductor, capacitor, and the power MOSFETs, and output the lower voltage with high efficiency.

The Chapter 6 covers the layout design of the power management CMOS ASIC which uses the TSMC 65 *nm* technology. The pads distribution with the anti-ESD design is also included. Then the test board is designed. At last, test of the ASIC, read out the data, and compare the post simulation result with the test data.

The Chapter 7 deals with the outlook and conclusion.

The Appendix A introduces the CGEM-IT front-end design, which is the application background of the power management. This chapter describes the CGEM-IT detector principle, and gives out the design details of the preamplifier and the shaper with the baseline holder.

The original contributions of this thesis are as following. The first one is the H shape transistor. This is one new structure which have some radiation-aware ability for the TID radiation. Comparing the ELT structure, the H shape structure have less anti-radiation ability, but it uses less area, and it can get more flexible size (CMOS ratio). It can also use the normal simulation model to extract parasitic parameters for the post simulation, and contrasting to this, the ELT need special simulation model. The H shape transistor is more suitable for the radiation-aware (light anti-radiation) design.

The second original contribution is the new structure bandgap without OPA. This kind of bandgap not only can work in very low supply voltage (low to 0.9 V), but also use very few current. Comparing the bandgap with OPA, all the current is used for the temperature compensation, this new structure has more current efficiency. Moreover, this kind structure uses the PMOS, instead of the parasitic triode, as the negative temperature coefficient generator. Comparing the classic triode, the PMOS has better anti-radiation ability. This kind of structure is more suitable for the advanced technology ASIC in high energy physics application.

Chapter 2

The Radiation-Aware Methods on CMOS Technology in High Energy Physics

In high energy physics applications, most of the electronics is exposed to the high energy photons and particles. Generally, there are two kinds of high energy physics application situations. One is to detect high energy particles in space such as the space station (AMS, HERD etc.) or the dedicated satellite (EXOSAT, BeppoSAX, DAMPE, HXMT etc); the other is to detect high energy particles at accelerators such as at LHC at CERN, BELLE at KEK, BESIII at IHEP etc. The electronic equipment will encounter different high energy photons and particles according to the different situations. The space radiation effect is shown in Figure 2.1.

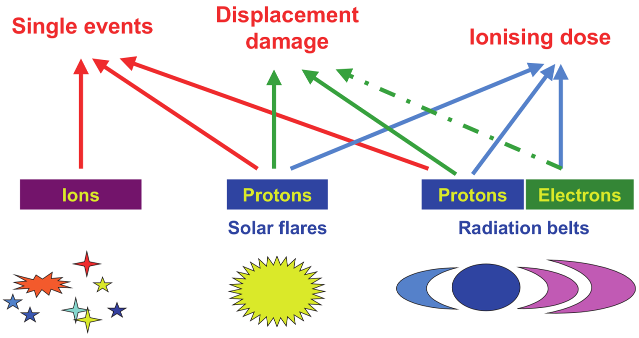


Fig. 2.1: Space radiation source and the possible damage effect

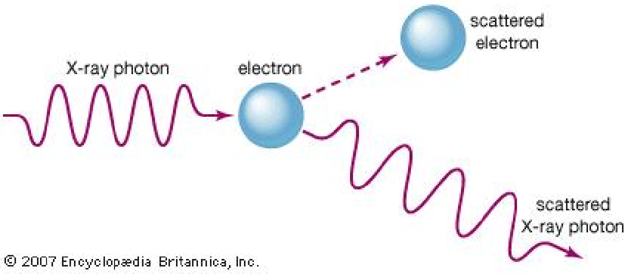
2.1 Interaction between High Energy Particles and Semiconductor Material

The nature of the radiation effects is due to high energy photons and particles interacting with the material such as the semiconductor, and changes the electron or atomic nucleus of the semiconductor, then changes the character of the semiconductor. So the first thing is to explain the interaction between the high energy radiation and the material.

The high energy photons, such as X-rays or gamma rays, can produce photoelectric effect, Compton scattering effect and pair production effect in the semiconductor material. And the high energy particles will mainly produce the ionization effect in the semiconductor material.

When the energy of the photons is about several keV, the main effect is the photoelectric effect. Some electron-hole pairs can be generated when a high energy photon is absorbed in the material. In the photoelectric effect the photon energy is fully absorbed by one of the electrons of the inner shells (K and L shells). Therefore the electron gets a large kinetic energy which will be lost crossing the material, causing the creation of electron-hole pairs.

When the energy of the photons is about several hundred keV (corresponding to an X-ray or gamma ray), the main effect is the Compton scattering effect. When a high energy photon interacts with a quasi-free electron (of the external shells) a Compton scattering occurs, as shown in Figure 2.2. Some energy is transferred from the photon to the electron and on the contrary the photon energy is lowered.



© 2007 Encyclopædia Britannica, Inc.

Fig. 2.2: Compton effect diagram

When the energy of the photons is above 1024 keV, the main effect is the pair production effect. If the energy of the photons is high enough, when the photons go through the nucleus, under the nuclear coulomb field, the high energy photon will create a pair of one positron and one negative electron. And then the positron stops in the material, and annihilates with an electron of the material, producing a photon pair. The secondary energy photons will continue to interact with the semiconductor, and thus form the showering, get a lot of low energy photons and electrons. Pair production effect is shown in Figure 2.3.

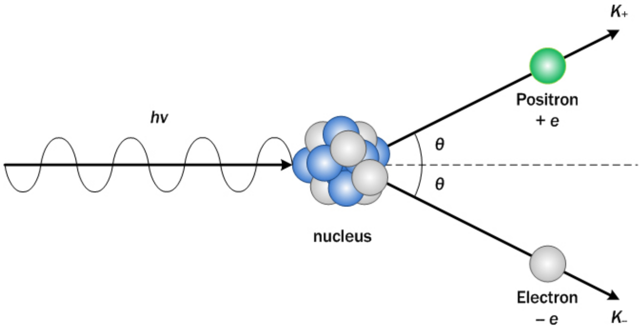


Fig. 2.3: Pair production effect diagram

All these three effects result in the production of one electron with a certain kinetic energy (depending on the effect and on the energy of the photon), which has been, partially or totally, transferred by the photon. Then this electron will deposit its energy in the medium. In a semiconductor material the deposited energy during the slowing down of the electron can produce a small or a large number of electron-hole pairs.

The high energy particles which can radiate the semiconductor are divided into two classes: charged particles and neutral particles, such as neutrons. When the neutron occurs the elastic collision with the nucleus in the semiconductor lattice, if the neutron kinetic energy is high enough, it will hit the nucleus out and make the lattice having some defect. The defect lattice will influence the character of the semiconductor.

The high energy charged particles are mainly electrons, protons, and heavy ions. The effect between the charged particles and the semiconductor is the electromagnetic interaction, including the ionization, hit out, coulomb scattering and so on. Among them, the main radiation effect is the ionization. When the incident charged particle interact with the orbital electrons, by means of Coulomb interaction, a loss of the energy occurs. At the same time, the orbiting electron get some energy.

When the orbiting electron gets enough energy to overcome the bond of the nucleus, the orbiting electron becomes a free electron. This is the ionization. The result of the ionization is positive ions and free electrons. If the inner shell electrons are thrown out, the outer electron will transfer into the inner shell for filling, at the same time an X-ray or an Auger electron is emitted. If the orbiting electron gets fewer energy and thus cannot overcome the bond of the nucleus, the orbiting electron will transfer to an higher energy level. This is the excitation of the atom. The atoms of the excited state are not stable, so after a short time, they will go back to the ground state. When this course happens, the released energy will be in the form of fluorescence.

To measure the ionization effect, the normal method is to measure the lost energy ($-dE/dx$) of particles through some material. The $-dE/dx$ can be expressed by the Bethe Bloch formula (2.1).

$$-\frac{dE}{dx} = \frac{4\pi}{m_e c^2} \cdot \frac{nz^2}{\beta^2} \cdot \left(\frac{e^2}{4\pi\varepsilon_0}\right)^2 \cdot \left[\ln\left(\frac{2m_e c^2 \beta^2}{I \cdot (1-\beta^2)}\right) - \beta^2\right] \tag{2.1}$$

where c is the speed of light, ε_0 is the vacuum permittivity, $\beta = v/c$ is the particle speed, e and m_e are the electron charge and rest mass respectively. Here, the electron density of the material can be calculated by

$$n = \frac{N_A \cdot Z \cdot \rho}{A \cdot M_\mu} \tag{2.2}$$

where ρ is the density of the material, Z is the atomic number, A is the relative atomic mass number, N_A is the Avogadro number and $M\mu$ is the Molar mass constant. The high energy charged particle ionizing the semiconductor is shown in Figure 2.4.

In addition, the main radiation damage caused by the high energy photons is also ionizing effect. Because these high energy photons (X rays or γ rays) will be converted into high energy electrons, protons, and ions in a very short time. These second charged particles will ionize the semiconductors and result in the main radiation damage. The different radiation effects caused by the particles and photons are shown in Figure 2.5. The more electron-hole (*e*-*h*) pairs means the more radiation damage.

From Figure 2.5, it can be found that the heavy ions will give the largest radiation damage. But the heavy ions are rare. The protons will be often a lot in these

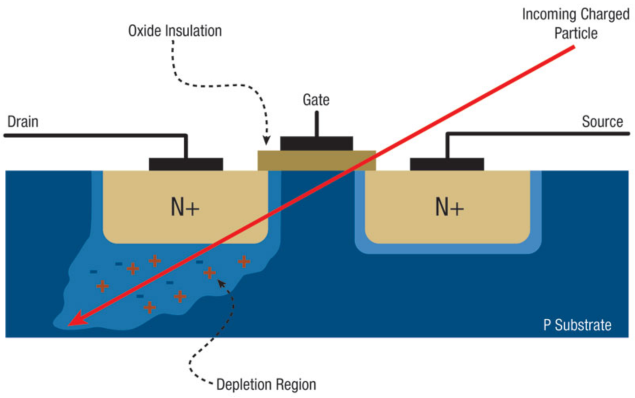


Fig. 2.4: Ionization effect diagram

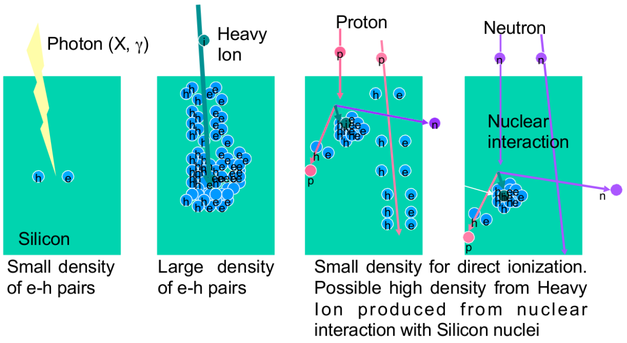

Photon (X, γ)
Heavy Ion
Proton
Neutron
Silicon
Nuclear interaction
Small density of e-h pairs
Large density of e-h pairs
Small density for direct ionization. Possible high density from Heavy Ion produced from nuclear interaction with Silicon nuclei

Fig. 2.5: Schematic view of the density of *e*-*h* pairs caused by different radiation

situations, and the proton can also create a large number of *e*-*h* pairs. So the protons often cause the main radiation damage.

2.2 Radiation Effects on CMOS Semiconductor

According to the radiation process, the radiation effect can be divided into two classes. One is the accumulation radiation effect, and the other is the non-accumulation radiation effect. The slow accumulation effect includes the ionization effect (Total Ionizing Dose, TID) and the non-ionization effect (Non-Ionizing Energy Loss, NIEL). The non-accumulation radiation effect is named the Single Event Effects (SEE). The TID and NIEL depend on the time, but the SEEs are random. The SEEs depend on the position of the hits, the energy, the ionization state of the particles and the working state of the circuit.

The NIELs are often caused by the high-speed neutrons and can kick the nucleus out of the lattice, therefore resulting in the displacement effects. The displacement effects often happen in the bipolar technologies and the optoelectronic devices. CMOS technology devices will not be influenced. The CMOS devices are mainly affected by the TID and the SEE. The radiation effects on the electron devices can be seen in Figure 2.6.

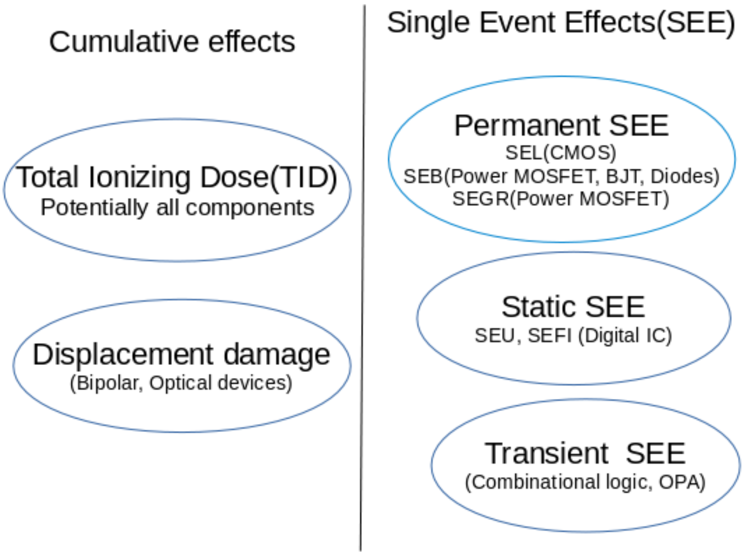


Fig. 2.6: Radiation effects on the semiconductor devices

2.2.1 TID Effects on CMOS Technology Devices

TID is the measurement of the total dose, which is the deposited energy in the material. The unit is *Gray* in the International System (IS), but the radiation effects community still often uses the old unit, *rad*(Si). One should get used to both of them, because the dosimeter persons speak about *Gray*, whilst electronic engineers working on the effects always speak about *rad*(Si). Luckily, the equivalence between the two is easy to count:

$$1\ Gray(Gy) = 100\ rad \tag{2.3}$$

In CMOS technology devices, the TID is mainly caused by the ionizing radiation effect. The high energy photons and charged particles will produce many electron-hole pairs in the CMOS device, especially in the oxide place. The TID mainly happen in two places. One is the gate oxide, and the other is the insulating oxide (or the field oxide). The insulating oxide is the LOCOS in the old technology and the STI in the recent technology.

Threshold Voltage Shifts

The TID effect course in the gate oxide is shown in Figure 2.7. The charged particles ionize the oxide of the gate, and then the electron-hole pairs are produced.

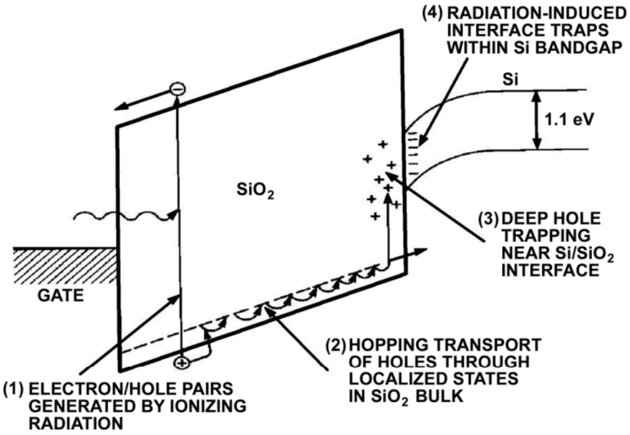

(4) RADIATION-INDUCED INTERFACE TRAPS WITHIN Si BANDGAP
Si
1.1 eV
SiO$_2$
(3) DEEP HOLE TRAPPING NEAR Si/SiO$_2$ INTERFACE
GATE
(2) HOPPING TRANSPORT OF HOLES THROUGH LOCALIZED STATES IN SiO$_2$ BULK
(1) ELECTRON/HOLE PAIRS GENERATED BY IONIZING RADIATION

Fig. 2.7: TID effect on the gate oxide

The ionized electrons will go to the gate side easily under the electric field, and at the end go out of the CMOS. In contrast, because of the ionized holes bigger size and slower moving speed, most of the holes will stay in the oxide. In absence of electrons, the hole cannot recombine and are trapped in the oxide. Therefore, the gate oxide will accumulate plenty of holes with much positive charge.

To the transistors, this positive charge will change the threshold voltage. To the NMOS transistors, the threshold voltage is positive, so the trapped positive will reduce the threshold voltage. To the PMOS transistors, the trapped positive holes will also reduce the threshold. But because the threshold voltage is negative, the absolute value increases. The changed threshold voltage V_{ox} is

$$\Delta V_{ox} = -\frac{q}{\mathrm{C}_{ox}}(\Delta N) \tag{2.4}$$

where q is the electron charge 1.6×10^{-19} C; C_{ox} is the specific gate capacitance, the unite is F/m^2; ΔN is the surface density of the trapped holes; the unit is m^{-2}.

Besides the trapped holes in the gate, the radiation also creates the interface states between the oxide and the crystalline silicon. The build-up course of the interface states is very slower than the direct hole trapping, so the device characteristics will change continue even if the irradiation action has stopped [1].

The polar of the interface states depends on the CMOS type, so it is amphoteric. The interface states can be the donors or the acceptors. A donor trap releases an electron when it passes from below to above the Fermi level, whereas an acceptor trap captures an electron when it passes from above to below the Fermi level. The interface oxide trapped charge is responsible for a variation ΔV_{in} in the threshold voltage: this can be positive or negative.

$$\Delta V_{in} = -\frac{\mathrm{Q}_{in}}{\mathrm{C}_{ox}} \tag{2.5}$$

where Q_{in} is the interface trapped charge, positive or negative, and C_{ox} is the specific gate capacitance, the unit is F/m^2.

If a transistor is radiated when the gate is biased to a certain working voltage, the gate electric field will depart the electron-hole produced by the radiation. If it is a NMOS, the gate working voltage is positive, so the holes will move towards the interface between the gate and the channel; the electrons will be attracted to the boundary between the gate and the electrode. If the transistor is a PMOS, the opposite is true.

In the NMOS transistor, the holes directly trapped in the gate oxide attract electrons in the channel and they favor the channel inversion, lowering the device threshold voltage. The interface states, on the other hand, trap electrons becoming negatively charged, therefore they act in the opposite direction. This is at the origin of the “rebound” phenomenon observed in irradiated transistors, where the threshold voltage first decreases and then somewhat recovers [1].

In PMOS transistors, the holes trapped in the oxide repel the ones in the channel, hence they hinder the channel inversion. The empty interface states are also positively charged and act in the same direction, therefore the effect of irradiation is to increase the absolute value threshold voltage [5]. Based on the above analysis, the whole voltage variation is:

$$\Delta V = \Delta V_{in} + \Delta V_{ox} = -\frac{q}{C_{ox}}[(\Delta N]) - \frac{Q_{in}}{C_{ox}} \tag{2.6}$$

where the first item is always negative, and the second item is negative to the PMOS, positive to the NMOS.

The threshold voltage variation is shown in Figure 2.8. Where ▲ represent the threshold voltage caused by the interface states trapping charge, and ■ represent the gate oxide trapping charge. As the radiation going on, the NMOS TID radiation can recover slowly, but the PMOS cannot recover by itself [6].

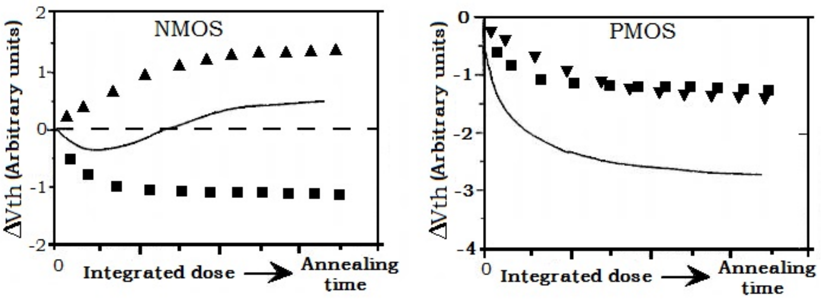


Fig. 2.8: Threshold voltage variation under the TID

For the threshold voltage shift caused by the TID effect, the NMOS and PMOS can be annealing once stopping the radiation and finally, the gate oxide holes can recover, but this course will not happen in the interface states.

There are two kinds of annealing: tunnel annealing and thermal annealing. The tunnel annealing is caused by the quantum tunneling effect. Quantum tunneling or tunneling refers to the quantum mechanical phenomenon where a particle tunnels through a barrier that it cannot surmount classically.

In the classical theory, the low energy electrons cannot go through the interface between the S_iO_2 and S_i which has a higher barrier, but according to the quantum theory, the low energy electron has some probability to tunneling the barrier. The electrons from the silicon can tunnel to and recombine with the trapped near S_iO_2 interface. It is shown in Figure 2.9.

As a consequence of the exponential decay of the tunneling probability with the depth in S_iO_2, at a given time t, the hole traps are emptying at a depth $X_m(t)$ from the silicon.

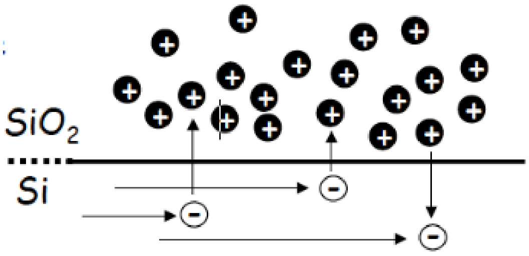


Fig. 2.9: Tunneling effect on the interface

$$\mathrm{X_m}(t) = \left(\frac{1}{2\beta}\right) \ln\left(\frac{t}{t_0}\right) \tag{2.7}$$

where β is the tunneling barrier height parameter, and t_0 is the time scale parameter. By applying a positive bias to the NMOS gate field, it have the effect of increasing the annealing rate. That means a positive electric field lowers the barrier for tunneling.

The second annealing effect is the thermal annealing. Electrons in $\mathrm{S_iO_2}$ can recombine with trapped holes; this process can be modeled with an emission probability $\mathrm{p_{em}}$ of an electron from the valence band of $\mathrm{S_iO_2}$ to the traps.

$$\mathrm{p_{em}}(\mathrm{T}) = \mathrm{AT}^2 \cdot \exp\left(-\frac{q\varphi}{k_B\mathrm{T}}\right) \tag{2.8}$$

where A is constant depending on trap cross section, T is absolute temperature, φ is the energy difference between the trap level and the valence band, q is the electric charge $1.6 \times 10^{-19}\ C$, and k_B is the Boltzmann constant.

Radiation Leakage Currents

As said before the TID radiation effect also produces the electron-hole pairs on the STI or LOCOS which are made of the oxide. The STI is the newer CMOS technology.

In CMOS technology, the channel length and the width are the two most important parameters. To get the precise size, the gate always covers the diffusion area and extents to the insulation area, which is the STI or the LOCOS.

The LOCOS can form the Bird's Beak as shown in Figure 2.10. The Bird's Beak not only occupies larger area but also produces the parasitic CMOS under the TID effect, as shown in Figure 2.11.

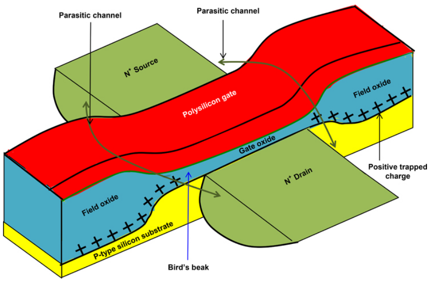


Fig. 2.10: Parasitic CMOS formed by Bird's beak

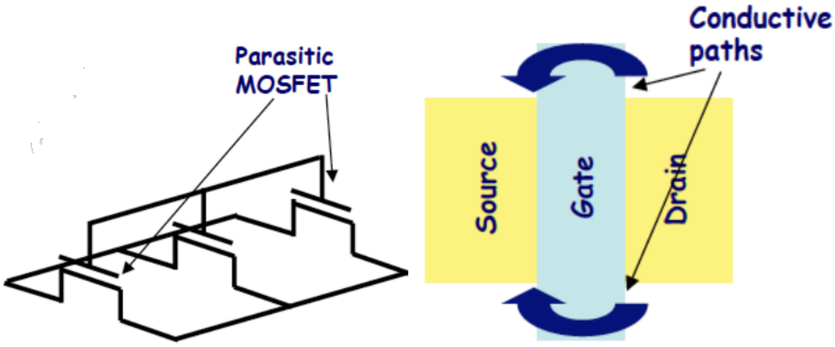


Fig. 2.11: Parasitic CMOS schematic

The radiation effect on the LOCOS oxide is the same as the gate oxide. The parasitic channels are between the source and drain terminals under the bird's beak region. In the NMOS transistor, the holes trapped under the bird's beak can invert the parasitic NMOS channel even if the main NMOS is off. And the parasitic transistor located at the channel boundaries will be conducted, then the leakage current appears.

Like the gate oxide, the interface states here also counteracts this course. So after a delay time, the function of the interface states will play a part. The leakage current increase at first, and then decrease after a certain time even at that time there are higher doses. At that time there is the sufficient number of interface states appearing. The parasitic transistors in parallel with the main device feature a different width W but the same length L.

In the PMOS transistors, the current carriers in the channel are the holes, so the holes created in the bird's beak by the TID cannot invert the channel, so the parasitic transistor cannot form. So there is not the leakage current in the PMOS. **So This effect exists only in NMOS transistors.**

In the newer STI technology insulation oxide, there is not the bird's beak, but also have the parasitic CMOS phenomenon under the TID radiation [7]. The STI architecture is shown in Figure 2.12.

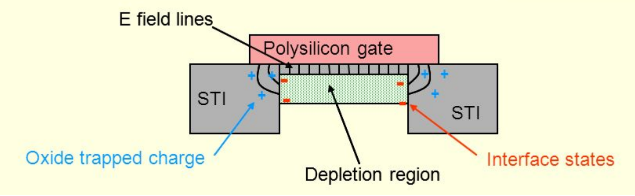


Fig. 2.12: STI architecture and the parasitic channel

The reason is that under the electric field of the gate voltage, the holes caused by the TID can accumulate on the steep slope of the STI, which instead of the bird's beak effect in the LOCOS.

The current caused by the parasitic transistor on both sides of the channel is just one kind of the leakage currents. Besides this, there are also other three kinds of leakage currents. They are the leakage current between the N diffuse of different NMOS, between the N different of the NMOS and the NWELL, and between different NWELLs respectively. The parasitic conductive path is formed between the diffusion areas of two different devices. It is shown in Figure 2.13.

In this Figure, the polysilicon here can play a role of the gate to absorb the electrons and leave the holes on the substrate. The polysilicon maybe come from some polysilicon resistors or some dummy transistor. The holes trapped in the oxide can form an inversion layer at the boundary between the STI and the substrate. At this time, the different adjacent N+ diffusion has different potential, then the parasitic current can flow.

This leakage current in the STI or LOCOS has the relationship with the thick of the oxide and the layout. So even using the same design, different process and technology will have the different leakage under the same TID radiation.

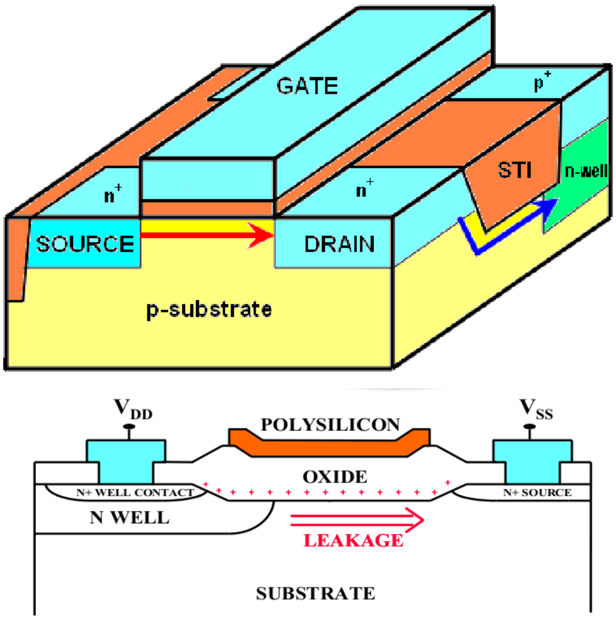


Fig. 2.13: Cross-sectional diagram indicating: (1) drain-to-source leakage and (2) leakage between the n+ source/drain region of an n-channel device and the n- well region of an adjacent p-channel device.

2.2.2 SEE Effects on CMOS Technology Devices

The single event effects (SEE) comes from the interaction of an individual particle with the circuit. There are many kinds of SEE effects, including SET (single event transient), SEU (single event latch-up), SED (single event disturb), SEFI (single event functional interrupt), SEGR (single event gate rupture), SEB (single event burnout) and so on.

Among them, some SEEs produce the soft errors which means these errors can be repaired, and some SEEs produce the hard errors, which means this kind of errors often result in the device damage.

The soft SEEs include the SEU, the corruption of a single bit in a memory array; the MBU, the corruption of multiple bits due to a single particle; the SET, an error signal induced by an ionizing particle.

The hard SEEs include the SEGR, rupture of gate oxide occurring especially in power MOSFETs; the SEB, the burnout of a power device; the SEL, the activation of parasitic bipolar architectures, leading to a sudden increase of the supply current.

In CMOS technology, the main SEE effects are the SEU/SET and the SEL.

SEE/SET Effect

The SEU/SET mainly results in the flipping of one or more bits in the digital registers. In general, this is a threshold phenomenon that depends on the Linear Energy Transfer (LET) of the impinging particle, because the deposited charge must be sufficient to change the status of the circuit node. The smaller the node capacitance, the more probable the upset. It is shown in Figure 2.14.

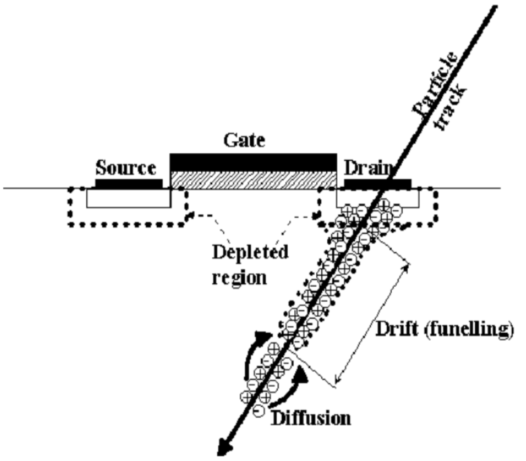


Fig. 2.14: Single Event Upset (corruption of logic state)

The LET threshold depends on the ratio between the node store charge and the charge resulted from the particle radiation. When using the advanced technology with the lower voltage and smaller node capacitance, the node charge will reduce, so it seems the bit will more sensitive to the particle radiation. But this is not the case. Some semiconductor companies have researched the SEU LET for a long time. They find, to a simple one bit, it is harder to flip over under the same particle radiation, and to the whole chip, it is easier to happen the SEU. It is easy to understand because the advanced technology has more bits than the old technology [5]. It proves that not only the node capacitance and the supply voltage, but also the sensitive area and the charge collection efficiently will play an important impact on the single particle radiation.

Recently, most of the semiconductors companies which do not do the radiation tolerant components also research the SET sensitive for their products. The reason is the products can suffer the neutron radiation which can play a part when using the advanced technology [8]. Most of the neutrons come from the cosmic rays, and the neutrons especially the thermal neutrons can raise the Failure In Time (FIT), reduce the MTBF, and therefore shorten the lifetime of the products.

The possible reason is that the thermal neutrons or the slow neutrons with around 25 *meV* have a large cross-section for interaction with an isotope of Boron (^{10}B). In the natural Boron, the Boron (^{10}B) occupy about 20%. So when a semi-conductor uses the boron doping, there a radiation effect could happen. Then ionizing particles such as the Li and Alpha particle can be produced, which can upset the digital node. It is shown in Figure 2.15. If this mechanism can be verified, the semiconductors will not use the Boron doping and raise the FIT of the components [9].

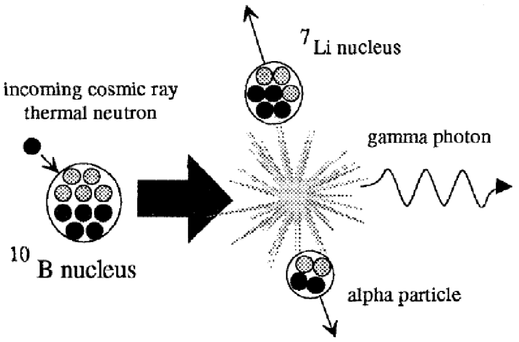


Fig. 2.15: Fission of (^{10}B) induced by the neutron

Another important upset mechanism appears in the advanced technology in recent years. It is the SET (Single Event Transient). When the particle hit the combinatorial logic, the induced pulse can propagate through the logic unit until reach a register where it can be latched. It is shown on the left picture in Figure 2.16. It can be seen that the signal flipping consist of two parts: one has no relationship with the frequency, which is the SEU; the other is linearly proportional to frequency, which is the SET.

In the low-frequency system, some SET errors can recover and disappear by themselves. But in the high-frequency system, the SET errors can be transferred to the register and be latched. So reducing the frequency is one method to lower the single particle flipping.

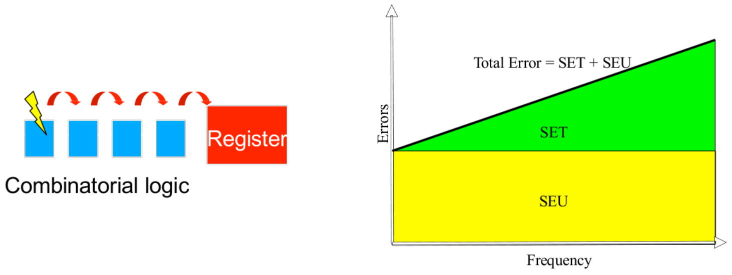


Fig. 2.16: SET source and the result

SEL Effect

The SET is the signal event latch. When the NMOS and the PMOS are near together in the P-substrate, there will be two parasitic triodes, which are shown in Figure 2.17.

The triode, another name is BJT (Bipolar Junction Transistor), is composed by two PN junctions which use the same common P terminal or the same common N terminal. In Figure 2.17, the P diffusion in the PMOS and the Nwell make up one PN junction, and the Nwell and the P-substrate make up another PN junction. Both of the two junctions use the same N terminal, so they compose the PNP triode which is the vertical PNP.

On the other hand, the N diffusion of the NMOS and the P-substrate make up one PN junction, the P-substrate and the Nwell make up another PN junction. Both of them use the same P terminal, so they compose the NPN triode which is the lateral NPM.

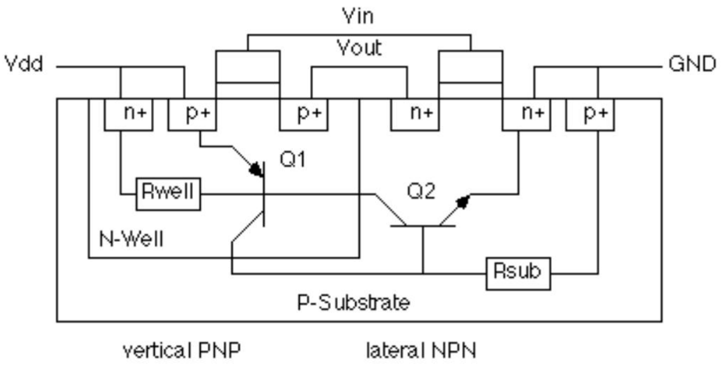


Fig. 2.17: Cross section of parasitic transistors

At the same time, normally the P-substrate and one of the N diffusion in the NWLL will connect to the GND; the Nwell and one of the P diffusion in the PMOS will connect to the VDD. Together with the parasitic resistor Rwell and Rsub, the Latch-up circuit will form which is shown in Figure 2.18. The Q1 and Q2 represent the PNP and NPN triodes respectively.

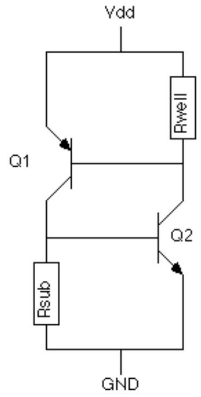


Fig. 2.18: Schematic of parasitic transistors

If the trigger current flows into the Rwell and Rsub, the two parasitic resistors will have the voltage drop. When the voltage drop can reach above 0.7 V, it can switch on the triodes, and then the two triodes can form the positive feedback. When the current flowing through the Psub, it raises the base voltage of Q2, the Rsub current will be enlarged by the NPN triode, so it can enlarge the Rwell current. After that, the current through the Rwell can switch on the Q1, at the same time, the Rwell current will be enlarged by the PNP triode. This enlarged PNP current again flows through the Psub and enlarges the NPN current. When the trigger current flows into the Pwell, the positive feedback is the same.

The trigger current not only results from the cosmic rays, the accelerator, and other high energy physics application surroundings but also comes from the static electricity, the interface I/Os or other disturb which can easily inject the current through the pads and the interface. The interface circuit will more easily cause the latch-up problem.

When the latch-up happens, because of the positive feedback, the latch-up is equivalent to create a short path from the power supply and the ground. The short circuit can burn the device at the end. When this situation happens, the best method is cutting off the power. And then after a few time, restart the device again.

From the beginning of CMOS technology up to now, the latch-up problems always trouble the designer. Although the STI, the retrograde wells, and the lower supply voltage of the advanced technology have released the latch-up problem, the latch-up cannot be immune. It also depends on the layout.

From what discussed above, if the latch-up happens, the following conditions must be satisfied. (1) both triode's conduction creates a low resistance path between VDD and GND. (2) the product of the gains of the two transistors in the feedback loop, b1 × b2, is greater than one. The b1, b2 here are the gains of the Q1, and b2 respectively.

2.3 Radiation-Aware Methods on CMOS Technology Devices

According to the radiation effects on CMOS technology devices, the main radiations are the TID, SEE/SET and the SEL. So the radiation-aware methods against these are introduced. Meanwhile, as the technology advanced, some of these methods are improved.

2.3.1 Radiation-Aware Methods for TID Effects

The TID effects on CMOS technology devices include the threshold voltage shifts, the leakage current, mobility degradation, and the sub-threshold slope change. The most influential are the threshold voltage shifts and the leakage current [10].

Radiation-Aware Method for Threshold Voltage Shifts

To the threshold voltage shifts caused by the TID effect, the ionized electron-hole pairs are produced when the high energy charged particle going through the gate oxide. So the thinner the gate oxide, the fewer electron-hole pairs produced. Meanwhile, the interface states also depend on the oxide thickness, and the advanced technology will get the fewer interface states effect [11].

At the same time, the thinner gate oxide, the easily the electron in the silicon can go through the interface between the oxide and the silicon, and then remove the trapped charge through the quantum tunneling effect. The more advanced technology can get the thinner gate oxide, and therefore get the fewer threshold voltage shifts. This has been confirmed since by several sources on commercial grade gate oxides of the advanced CMOS in Figure 2.19.

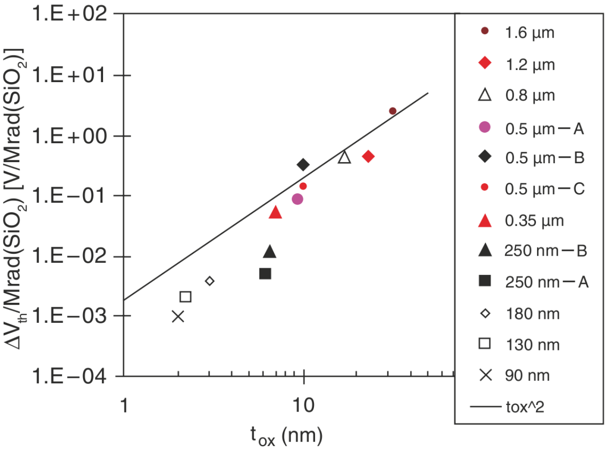


Fig. 2.19: Threshold voltage shift in different CMOS technologies

In this figure, the different technology nodes cause the different threshold voltage shifts. The test is under the same 1 *Mrad*(Si) of TID on the NMOS transistor samples.

When 1 *Mrad*(Si) is applied to the NMOS, it can be seen that at the 180 *nm* or below, the t_{OX} is less than 4 *nm*, and the threshold voltage shift V_{th} is less than 10 *mV*. The 1 *Mrad*(Si) TID is suitable for many space applications.

When using the 130 *nm* or 90 *nm*, the gate oxide thickness is about 2 *nm*, and the threshold voltage shift just 1 *mV* to 3 *mV*. It is at the same level with the statistical strategy error. When using the 2 *nm* thick oxide, the dose of the order 10 *Mrad*(Si) also can be ignored.

From what discussed above, the threshold voltage shifts mainly depend on the technology node. Fortunately, the current mainstream CMOS technology used in the front-end of high energy physics applications is less than the 180 *nm*. When used in the normal radiation surrounding, the threshold voltage shifts can be negligible.

Radiation-Aware Method for Leakage Currents

While the thickness of the gate oxide decreases, the STI or the LOCOS oxide does not change a lot. So the radiation effect on the STI or LOCOS does not reduce much. A significant amount of trapped holes and interface states are formed in the thick oxide.

As said before, there are four kinds of the leakage current caused by the TID, one is in the parasitic NMOS at the edge of the NMOS, and the three other kinds of leakage current (NMOS to NMOS; NMOS to NWELL; NWELL to NWELL) are the current between two different N+ diffusion.

For reducing the leakage current, the key point is to cut the possible conductive paths between drain and source of the same transistor and between the diffusion area belonging to different devices. This can be achieved by employing Enclosed Layout Transistors (ELT) and guard rings.

The ELT is edge-less transistors using the annular gate shape. This geometry reduces leakage current due to cumulative effects in NMOS, even at very high total doses, which is on the cost of a larger area. The PMOS will not generate the parasitic transistor and will not result in the leakage current. Therefore, PMOS do not require ELT shape. The ELT transistor is shown in Figure 2.20 which gives the difference between a standard device and an ELT transistor.

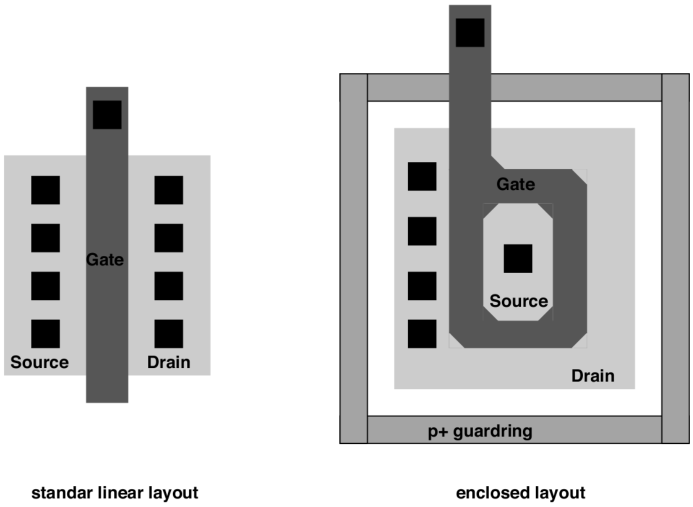


Fig. 2.20: Standard transistor (left) and the typical enclosed layout transistor (right)

From the figure, it can be seen that the gate of the ELT fully encloses one terminal of the electrodes. The transistor thus becomes asymmetrical, because the external electrode is larger than the internal one.

This asymmetry introduces a difference in the output resistance, so always the inner electrode is the drain, not only because of the inner electrodes having less parasitic capacitance, but also the less resistance. Figure 2.21 shows the radiation-aware effect by the ELT transistor.

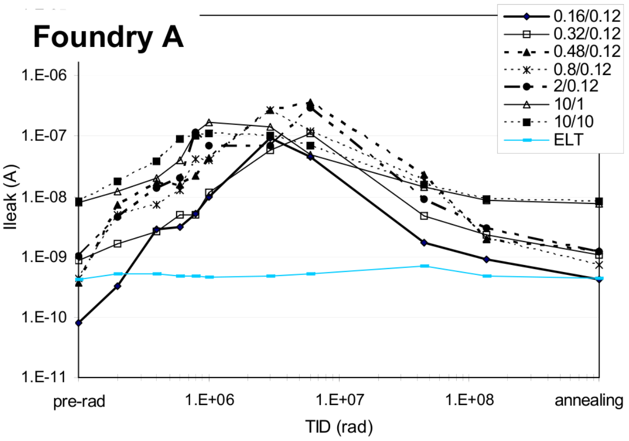


Fig. 2.21: 130 *nm* technology transistor leakage current under the different radiations

From the figure, the ELT can really reduce the leakage current. The annular gate prevents any conductive path between drain and source, thereby solving the issue of device level leakage. To cut possible parasitic paths towards neighboring devices, a ring of P+ substrate contact (P+ guard ring) is needed to surround the transistor.

The leakage current of the standard transistors have the common characteristics that the leakage current raises up as the more radiation and decline down as the radiation further growth. This can be explained by the interface state which plays a role after a certain time.

The use of ELTs allows us to implement circuits with extremely good radiation hardness comparing the standard CMOS technologies. Using the ELT for the designs, the ASICs tolerating more than 100 *Mrad*(Si) of total ionizing dose have been reported. Although it greatly enhances the TID tolerance, the adoption of ELTs in a design has also several drawbacks:

- Enclosed layout geometries occupy more area;

- Calculating the effective aspect ratio of an ELT is not trivial and the complicated formulas must be used;
- Matching between ELT devices is somewhat worse than standard ones;
- It cannot get the small aspect ratio.

Based on this drawback, this paper offers a new radiation-aware structure, the H shape layout, which is shown in Figure 2.22. In this kind of transistors, the gate oxide forms the H shape. The S and D terminals are on the two sides, which are half surrounded by gate oxide. When use the simulation software to post simulate the H shape transistors, the width is a litter larger than the designed oxide width. This difference comes from the wider gate oxide on the two terminals. The Cadence software has not such shape transistors simulation model. When using the H shape transistors to replace the standard transistors, it needs to change the H shape to suitable size in order to make the post simulation width and length to be almost the same with the standard transistors size.

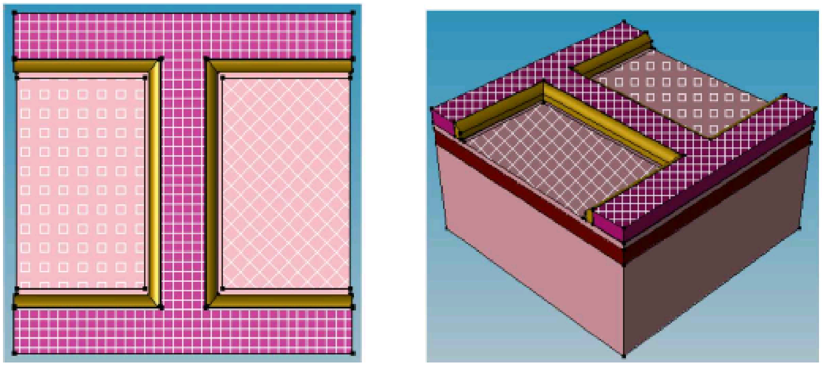


Fig. 2.22: A new kind of radiation-aware structure

The H shape structure combined with the advantages of the linear and ELT structures. Though not so thoroughly like the ELT, the H shape structure transistors cut the main leakage current road at the edge of the transistors. But a little current will exit at the two sides of the transistors. So the radiation-aware ability of the H shape transistors maybe not so strong as the ELT, but in most application, they are enough to use.

The advantage is the symmetry and small aspect ratio. So it can match better than the ELT and have the small aspect ratio. In addition, this kind of structure occupies less area than the ELT transistors. Thus the H shape structure transistors are more suitable for the compact radiation-aware CMOS design than the ELT.

To the leakage current between the different N+ diffusion of the two neighboring devices, the best method is using the guard ring. The guard ring can reduce the leakage current, which is shown in Figure 2.23.

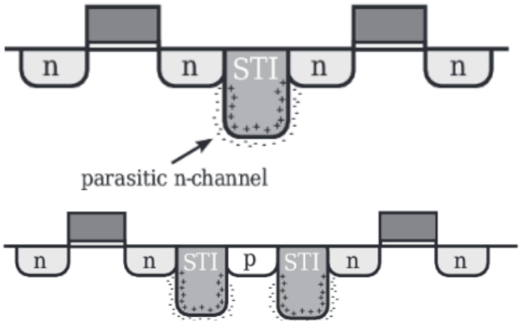


Fig. 2.23: Guard ring effects cutting the leakage current of the different N+ diffusion parts

In the P+ substrate, the P+ guard ring is needed to surround the N+ diffusion. The reduced the leakage current by the guard ring can be seen in Figure 2.24. The X-axis represents the TID radiation, the Y axis represents the leakage between the Nwell containing the PMOS transistors and the NMOS transistors. The NoGuard line means the Nwell do not use the guard ring to reduce the leakage current; the FullGuard line means the Nwell has a full guard ring, and the PartialGuard means the Nwell has the guard ring with a broken for the direct polysilicon connection between the NMOS gate and PMOS transistors.

From the result, the guard ring really reduces the leakage current even under the 10 $Mrad$(Si). When no guard ring, the leakage current ascend at first and descend later. It can be explained that the interface state have a role after a certain time [12].

The second method is that it should forbid the polysilicon (such as the polysilicon resistors) with the voltage between the neighboring N+ diffusion devices, in order to not separate the electron-hole pairs caused by the radiation.

2.3.2 Radiation-Aware Methods for SEU/SEL Effects

The main SEE effects on the CMOS technology devices are the SEU/SET and the SEL. These SEU/SET effect mainly happens in digital circuits, and the SEL effect

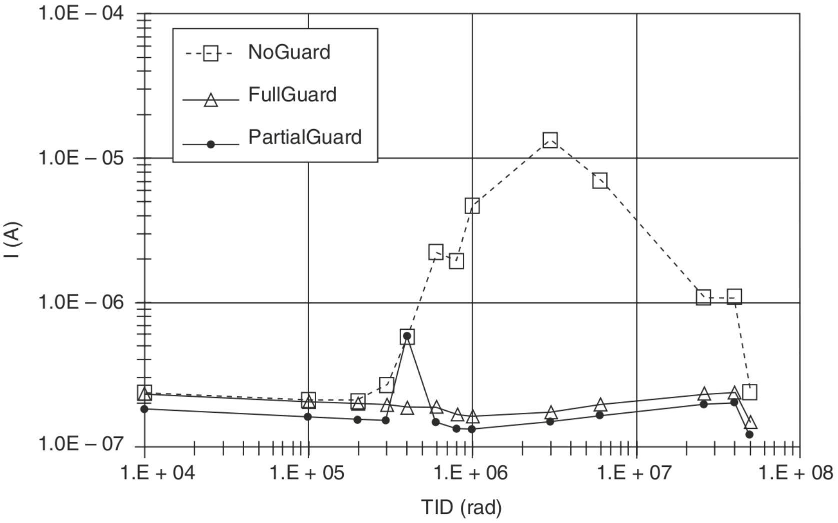


Fig. 2.24: Guard ring anti-leakage current effects

mainly happens in both of the analogue and digital circuits. If the front-ends of high energy physics applications are just analogue circuits, such as the preamplifier, the shaper or the power management circuits. The anti-SEL methods are enough for the design. If front-ends are mix-circuits, the anti-SEU/SET methods must be considered.

Radiation-Aware Method for the SEL Effect

The SEL effect is caused by the latch up in CMOS technology devices. Based on the ELT effect theory, there are the two methods to prevent the latch up effect. One method is based on the transistor level design, and the other is based on the systems level design.

The transistor level design includes two ways: the one is reducing the gain product b1 × b2 (reference Figure 2.18, the b1, b2 here are the gains of the Q1, and b2 respectively.), and the other method is reducing the nwell and substrate resistance (Rwell and Rsub respectively), producing lower voltage drops.

To reduce the gain product, there are two methods. One is increasing the doping density of the base terminal in the parasitic triode, and the other is enlarging the base terminal area. The triode work principle is: when the base terminal are positive enough relative to the emitter, the emitter will emit a lot of electrons to the base terminal. At the same time the base terminal will also send the holes to the emitter, but because the holes number is a little and the speed is slow, this part could be ignored. Since the base terminal is lightly doped and the thickness is very thin, so a few of holes will combine the electrons from emitter. Most of the electrons will flow into the collector terminal in the lift time by the inertial effect. So the key point is to reduce the gain. There are two ways, one is to enlarge the doping of the base terminal, which can recombine the electrons from emitter more easily. The other is to enlarge the base terminal area, which makes the electrons from emitter to the collector more difficult.

Based on the triode work principle, there are two tricks to reduce the gains. One trick is making the n+ layer buried in well to reduce the gain of Q1, and the other is moving n-well and n+ source-drain further apart in order to increases the width of the base of Q2. Fewer $\beta 2$ gain also reduces the circuit density. The STI technology could reduce the gains, because in the STI the thickness between substrate and the nwell is thicker than the non STI technology. It is shown in Figure 2.25. The figure shows that because of the STI technology, the base terminal of the parasitic triode is larger than the LOCOS, so it can reduce the electrons going through.

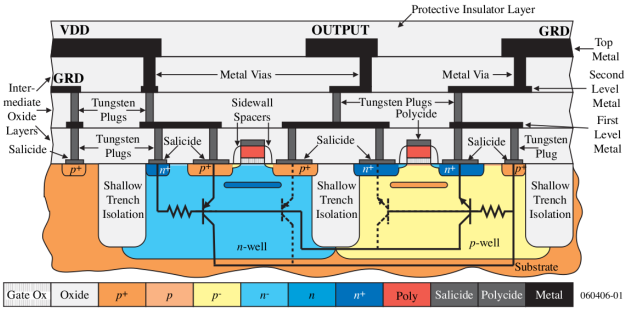


Fig. 2.25: Structure of the Shallow Trench Isolation

Though the guard ring cannot reduce the base terminal area, but it can boost the doping density. It is shown in Figure 2.26. Because the guard rings have the heavier doped than the nwell or the substrate, they could easily absorb a lot of electrons from the emitter. To some point, the guard rings are likely the collectors. So the n+ guard rings in nwell and p+ guard rings can reduce the gain of the parasitic triode.

To reduce the well and substrate resistances, one skill is reducing the R_{sub} by making low resistance contact to GND, and reducing the R_{well} by making low resistance contact to VDD. The specific operation is making nwell connectors and the substrate connectors as many as possible. The second skill is making the guard rings

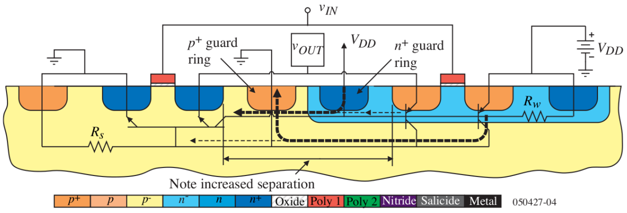


Fig. 2.26: Reducing the sensitivity to Latch up by guard rings

around p- and/or n-well, with frequent contacts to the rings, reduces the parasitic resistances. The third one is using higher substrate doping level reduces R_{sub}, such as the epitaxial wafers technology which can be seen in Figure 2.27.

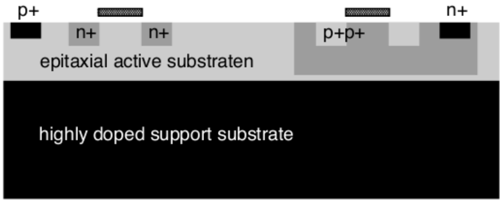


Fig. 2.27: Epitaxial wafers profile

The epitaxial wafer technology is that on the substrate, which is the lightly doped, there is the epitaxial layer which is heavily doped. The typical resistivity of the lightly doped wafer is about 20-50 $\Omega \cdot cm$ and the heavily doped wafer resistivity is about 20 $\Omega \cdot cm$. The resistors will be built on the heavily doped wafer which provides the low impedance path which can reduce the parasitic resistance. So it can reduce the sensitivity for the latch-up. At the same time, because of the reduced parasitic resistance, the nwell connectors and the substrate connectors are not needed so much.

The other radiation-aware method is the systems level design. The first item is to prevent the "hot plug in and out". It means that when the power supply is switch on, do not plug in or out some device on the board. The "hot plug in and out" can result in the large current on the pads of the components which can trigger the latch-up.

The second item is the I/O pads must be protected by the anti-ESD protection methods, which can clamp the pads voltage to the power supply or ground, and all the transistors connecting to the I/O pads must have the guard ring.

The third one is the circuits especially the digital circuits should avoid any transistors switching at the same time, which can result in large current going into or out of the chip. The sudden large current can easily trigger the latch up.

Radiation-Aware Method for the SEU/SET effect

The SEU or SET effect mainly happen in the digital circuits, which results in the flipping of one or more bits in the registers or the memory cells. Some of the SEU/SET have less harm, but some others, which can influence the configuration pattern or the states of a Finite State Machine (FSM) controlling the key operation, must be dealt with carefully.

The Hamming encoder, CRC32, or other kinds of encoding technology can be used for transferring the FSM. These special encode technologies can find the transfer errors and it can repair one or two error bits by themselves. These encoding methods are realized on the cost of the higher power consumption and more area.

The more common anti-SEU/SET technology is the Triple Module Redundancy (TMR). It is usually used in the crucial registers such as the SRAM configuring the FPGA. The TMR architecture is shown in Figure 2.28.

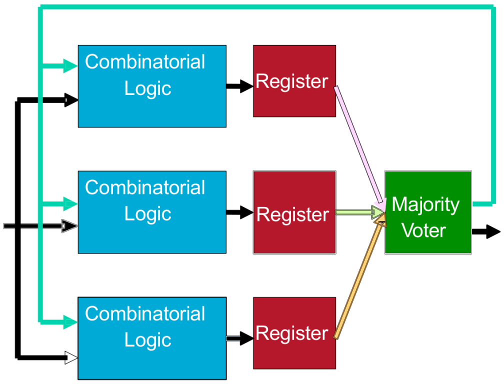


Fig. 2.28: Triple Module Redundancy (TMR)

The Triple Module Redundancy uses one or more voters, and the more voters has the better effect but needs more additional transistors. Normally, one voter is enough. To reduce the SET, it can use less frequency clock and also the time redundancy with TMR.

The time redundancy technology (shown in Figure 2.29) can reduce the SET error, but the timing constraints delay must be taken into the account. The Majority voter must wait a certain time to receive the three registers information. If some SET happens, after a certain time, the SET can recover, so this time redundancy technology can reduce some SET errors.

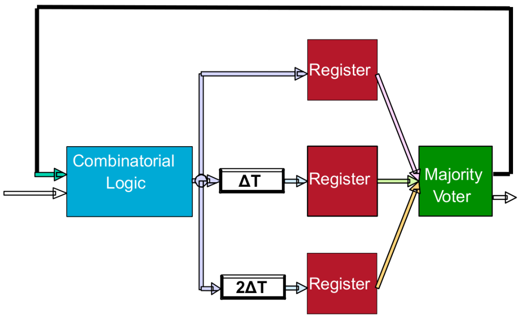


Fig. 2.29: Time redundancy with TMR

In the process technology level, to reduce the SEU effect, it can use the special process technology, such as the epitaxial substrates, the SOI and so on. The SOI introduces a kind of insulator onto the surface of the silicon substrate such as the sapphire, silica and so on. Generally, the insulator is the silica (S_iO_2), which is shown in Figure 2.30. But the SOI technology cannot reduce the TID effect.

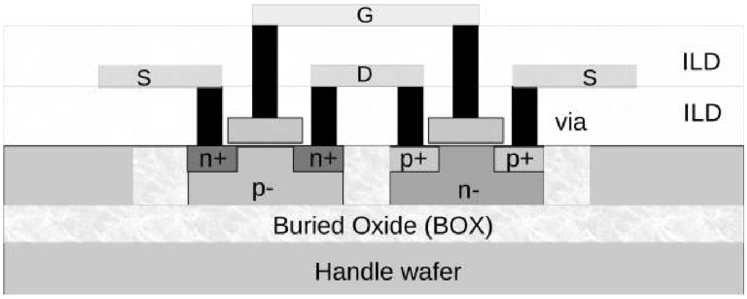


Fig. 2.30: Cross view of the SOI structure

The SOI makes every transistor insulate from each other, and makes an end to the SEE including the SEL, SEE, SET and so on in principle.

On the transistor level, increasing the critical node charge by the larger node capacitance is also one design method. This way can be easily realized by using the larger transistors or adding the "extra" metal to metal capacitors onto the nodes, which can make the nodes more difficult to set up. Besides these, there are also some special architectures which can be useful to the SEU/SET effect.

Because most of the SEU/SETs are soft errors. When SEU or SET happens, the last method is to restart the circuit. The SEU and SET will disappear after a short no power supply time.

Chapter 3

Bias and Power Management Issues in Front-End ASICs

3.1 Power Management in Front-End ASICs

In general, the power management is used for disposing of the power and ground system, which also includes the voltage and current supply. The power management includes the switch DC conversion, low drop out voltage, voltage reference, voltage and current bias, the current source, high voltage supply, the power grid distribution and so on.

Power management is one of the important modules which can affect the performance of front-end ASICs. According to different front-end features, there are different power managements.

From the previous chapters, the full front-end includes not only the analogue circuits (such as the preamplifier and the shaper), but also some digital circuits (such as LVDS transfer cells). Even some analogue digital mixed circuits may be included, such as ADC, DAC, TDC and so on. Whatever included in all cases, the analogue circuits are needed.

To digital circuits, power management requirements are not so critical. Digital circuits will work well as long as the power management can output enough current, and the ripples have less effect to the digital state.

To the analogue circuits and the mixed circuits, the power management is very important. For example, if the power supply for ADC is not clean, the resolution of ADC will be weak. The preamplifier is also sensitive to power supply. Power management can affect the energy noise and the time jitter. To reduce the effect of the analogue circuits from the digital, the analogue circuits and the digital circuits should be separated.

Normally, power management of front-end ASICs is divided into three parts. The first one is the power supply, which can offer enough power to the core and the other circuits. The second one is the bias design, including the voltage and current bias circuits. These bias circuits are used for setting the ASICs work state. The third one is the auxiliary power management, including the power and ground line distribution, the power network, the IR problem design and so on.

According to different places, there are two kinds of power managements: on-chip power management and off-chip power management. In early time, not only the power supply, but also the bias circuits come from the off-chip, such as the early front-end products of IDEAS company. Recently, as the integration level is higher, some power managements have been moved onto the chip.

When using the off-chip power management, the commercial power management components are normally selected. When purchasing this kind of components, the quality grade and the temperature range must be considered according to the different applications. For example in environments with very low temperature, or very high radiation surroundings, the power management should be suitable.

In front-end ASICs of high energy physics, the bias and power management are playing more important roles. As the technology progress, they will have more influence to key parameters. As high energy physics detectors become smaller and more density, such as the silicon micro strip detectors and silicon pixel detectors, the discrete circuits are not suitable. So the on-chip CMOS front-end ASICs become the mainstream technology, which have fewer volume, fewer parasitic parameters and lower power consumption. As CMOS technology developing, the cost is also decreased rapidly.

Desiring less power consumption and less volume, front-end ASICs need integrate more circuits, including ADC, DAC, LVDS, the power supply and so on. So nowadays, the mixed circuits front-end ASICs become the mainstream.

Under Moore's Law, the number of transistors per square inch on integrated circuits (IC) becomes twice every 18 months, which is shown in Figure 3.1. Up to 2016, the digital techniques have got to the 18 *nm* and the analogue techniques have got to the 40 *nm* [13]. The space between the transistors is smaller. In this situation, the digital circuits are easy to influence the analogue circuits through the same power supply and the substrate, so it needs better power supply including the current and voltage bias.

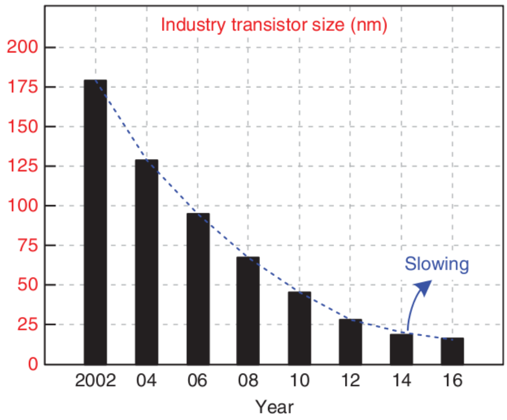


Fig. 3.1: Moore's law in recently years

At the same time, along with the less channel length, the electric field strength between the drain and source will become larger. If the power supply does not decrease synchronized, transistors will be breakdown by the strong electric field. So to protect the transistors, the power supply must decrease synchronized as the proceeding advanced [13], as shown in Figure 3.2.

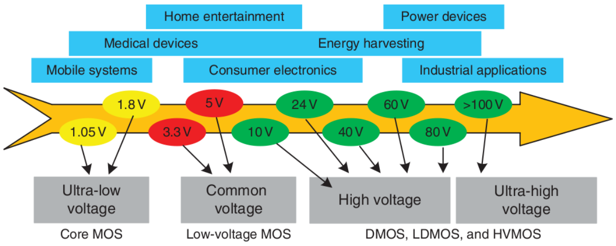


Fig. 3.2: Power supply variation as the advanced process

From the above figure, it can be seen that core CMOS transistor power supply has decreased to about 1 V. This situation gives a big challenge to design good power supply, bandgap, voltage bias, current bias and so on.

Although the advanced techniques are good for digital circuits, it has some drawbacks for analogue circuits. As the power supply decreases, the dynamic range becomes smaller, which will result in lower gain, and give some difficult for low noise. Considering the cost and performance, front-end ASICs in high energy physics often use the technology from 65 *nm* to 180 *nm* in these days. For example, BESIII CGEM-IT front-end ASIC named TIGER uses 110 *nm* UMC technology, and the CERN LHC third generation pixels detector project uses 65 *nm* TSMC technology. The both projects use 1.2 *V* as the power supply.

There are some challenges of the power management design in deep sub-micron technology. To the bandgap, the challenge is that, because of the low-power supply, the output of bandgap should be less than 1 *V*. So the classical architecture, which gives out the 1.25 *V* solid voltage, cannot be used. It must use the low voltage structure, which can give out the low voltage.

For the LDO, the challenge is that the intrinsic gain of short channel transistor will be lower, and the error amplifier will get less gain. So the LDO output will have more deviation from the desired voltage. When adding some cascade to boost the gain, it will introduce more poles to the error amplifier and it can result in the LDO system unstable.

To the switch DC-DC converter, the problem, which exists in the LDO, also present, because it also needs one error amplifier. The output amplitude of one ramp generator is limited by the power supply, so when designing the compensation network, the capacitor and resistor will be large, and the input voltage will be also limited.

As gate oxide becomes thinner, gate leakage current becomes larger by the quantum tunneling effect. So how to control the quiescent current is the large problem in all advanced technology.

Power management includes not only the power supply circuits which give good power to the analogue and digital parts, but also the current and voltage bias which is often be ignored. The power management also includes the auxiliary power circuits, such as on-chip power and ground distribution grids, bonding wires, pads distribution, potentials of the substrate and nwell connectors.

3.2 Current and Voltage Biases

The current and voltage biases are key parts, which makes front-end circuits have a good performance. But they are often be overlooked in front-ends especially the muti-channel systems.

They are always thought simply, so they are often left behind to be designed. But the bias voltage and current often introduce some problems, such as the limited area, the introducing noise, the PSRR. When these problems appearing, there are often no enough time and area to solve them. For example, when designing the BESIII CGEM-IT detector front-end CMOS ASIC (the TIGER first version), the bias current circuits introduced much noise into the preamplifier. But there was not enough time to optimize it, at the end added a lot of Ncap capacitors to solve it, which needed more area.

On the other hand, power consumption and area are often not considered by designers from beginning, so designers always leave small margin to the power and area. These situations make the bias circuits limited by area and power consumption, which make the design difficult.

The main aim of bias circuits is to adjust the circuits onto a right DC working points. There are two kinds of bias circuits: the voltage bias and the current bias. The voltage bias is always produced on the current bias, so designing a fine current bias circuit is the first key point.

3.2.1 Current Bias Design in Multichannel Front-Ends

Most front-end circuits in high energy physics are multichannel systems. In this situation, the ideal solution is that every channel has it own bias cells, but in fact they are always limited by the area and power consumption, so normally it needs some channels sharing one same bias voltage and current. A simple bias block [1] as shown in Figure 3.3.

In the simple bias schematic, R_{BIAS} series connected to a transistor is used for producing the suitable current, then this current is transferred to every channel through current mirrors. But this simple current bias has some problems.

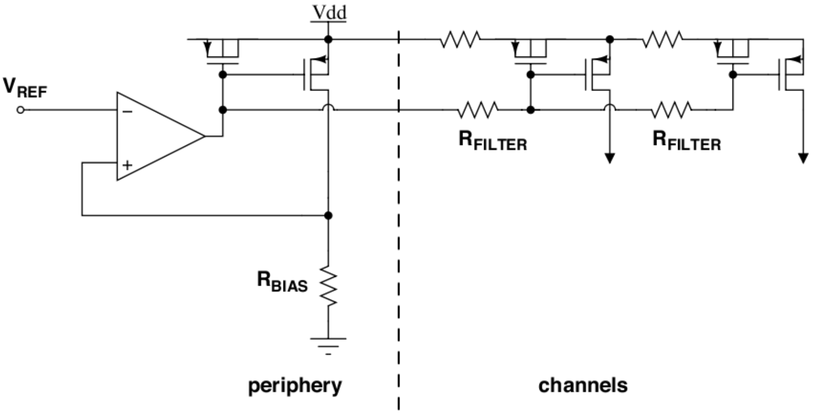


Fig. 3.3: A simple current bias schematic

The first problem is that the bias current is easily affected by the power supply, so the PSRR is very poor. In the mixed circuits, the power supply lines are often influenced by digital circuits. This kind of disturbance can transfer the noise into the bias current circuits, and then further transfer into all the channels.

The second drawback is the IR drop problem. Because the power and ground lines have some parasitic resistance. When the current flowing through the power and the ground lines, there must be some voltage drop along the lines, this voltage drop is called IR drop. The channels, which are nearer the bias circuits, will have fewer IR drops, and the channels which are farther will have larger IR drops. So it will make the channels have different relative bias voltage, increasing the channel inconsistency. The problem will be more serious when there is only one bias circuit, especially all the channels at one side of the bias circuit.

The third one comes from the common gate line, which is common to all channels, so there will be cross-talk when one channel disturbs the gate line. Especially when the PMOS is large, parasitic capacitance between the gate and drain will be large (sometimes dozens of pF), then the drain voltage will influence the common gate line through the parasitic capacitance. This situation can easily happen on the input transistor of front-end circuits. To release these problems, the book [1] gives out some solutions.

The merit of voltage transferring bias in multichannel system is that it can save routing lines, and the drawback is the large IR drop. Contrasting with this, the current transferring bias has the advantage of less IR drop problem but has more complex routing. The following circuit integrates the both advantages.

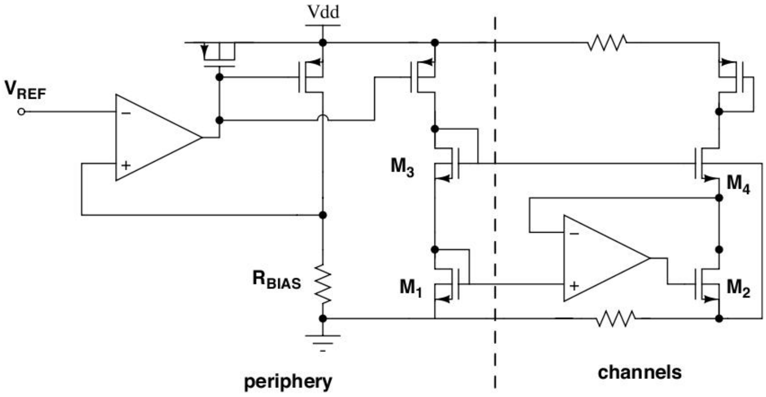


Fig. 3.4: An excellent current bias schematic[1]

In Figure 3.4, M_1 and M_3 form the periphery cascode architecture. And M_2 and M_4 form the channel cascode. The two cascodes form the cascode current mirror. The gate voltages of M_3 and M_4 are almost the same. Because in the gate line there is almost no current, voltage drop can be ignored. At the same time, the source voltages of M_3 and M_4 are also almost the same, because of the operational amplifier of the channel and the negative feedback circuit. So the currents of M_3 and M_4 can be determined by the aspect ratio. To get a precise ratio, the lengths of M_3 and M_4 are normally set the same, so the current ratio just depends on the width ratio [1].

Because this bias architecture transfers the voltage through gate line, one gate line can transfer to multichannel. But the operation amplifier also occupies some area, so a compromise method is the receiving circuit can be shared by the nearby channels. In this situation, even there are some IR drops, they can be ignored, because the sharing channels are very close. For example, if one such receiving current circuit can share the current bias with four channels each side. A 64 channel front-end just needs seven receiving circuits to finish the bias current.

From what discussed above, to get a better suitable current bias, the trade-off between the complex, the area, and the current accuracy must be considered case by case.

At the end, the quantum tunneling effect must be considered especially in advanced deep sub-micron technology. The tunneling effect can enlarge the gate leakage current of transistors. It can reach up to *pA* level and vary as temperature. One coin has two sides. When designing the very small current bias, this leakage current may influence the bias current. Sometimes the leakage current can reduce the impedance of the gate and finally influence the pole points. On the other side, when the *pA* level current needs be designed, it is hard to use the normal method to realize it. In this situation, it can use the gate leakage current to realize it. Considering the variation with temperature, enough margin must be left. One MOS capacitors is a fine device to realize the gate leakage current, because it can control the area of gate area easily.

3.2.2 Voltage Bias Design in Multichannel Front-Ends

The current bias has been discussed, the next important one is the voltage bias. Voltage bias is normally to bias the gate voltage of transistors. And voltage bias is generated based on the bias current.

The gate voltage of transistor is used to set the DC operating points, and to be sure the transistors working in the right state (such as the saturated state, the subthreshold state, the linear state, etc). And some cascode gate voltage also determines other transistors state. So voltage bias must be designed carefully and the voltage work points should be robust.

At the same time, the transistor work state and gate points can be affect by their parameters, but the parameter may change a few as the temperature variation and different chips. So the voltage bias should have the same trend with biased transistors. In order to realize this purpose, the bias transistor should be the same polar with the biased transistors. That means using the NMOS transistors to bias the NMOS transistors, and using the PMOS transistors to bias the PMOS transistors.

The voltage bias circuits are normally not the core circuits, so the designer often leaves this part a few power budget and area. Normally one voltage bias circuit can give out different bias voltages. When the voltage bias used in the multichannel circuit, one bias voltage bias may be shared by several channels.

Since bias circuits only support other analogue circuits by providing their DC voltages, the designer normally wants to minimize the bias circuit area and power consumption. Hence, one bias circuit is usually shared by several sub-circuits. An important practical consideration is, therefore, how to distribute the bias circuit across the chip effectively, without subjecting them to noise and inaccuracies caused by the mismatches.

Most of the bias circuit's outputs will be the gate voltages, which are used to bias current sources or bias the cascode transistors. When distributing these bias voltages, the devices, which are required to operate as current mirrors, should be placed in the same sub-circuit in near physical proximity, in order to ensure good matching between those devices. To reduce the noise, at the gates there are often some capacitors to filter the high-frequency noise.

3.3 Power Supply in Front-End ASICs

Normally the front-end electronic modules have the long-distance from energy supply. In order to save energy, the long-distance transmission voltage is high, such as the 12 V, 36 V or even higher. The dissipation power P_{diss} of the power line is as follows.

$$P_{diss} = \left(\frac{P_{rate}}{V}\right)^2 \cdot R_{line} \tag{3.1}$$

where P_{rate} is the request power dissipation by front-end modules, V is the transmission line voltage, and R_{line} is the transmission line resistance. To the same power line, the higher voltage, the lower dissipation on the transmission line.

In order to supply the front-end PCB board, there are two kinds of supplies. One is the low voltage for components and ASICs on board; the other is the high voltage for the detectors. Some detector even needs high to several thousands volts.

High voltage circuits are complicated and demands a higher reliability. The feature is high voltage but very few current. The current is normally no more than 1 mA. The normal design method is the charge pump, which uses switch capacitors to store and change the energy. Because of the complex, high voltage components are normally designed by the high-reliability company. And the high voltage has no relationship with front-end ASICs, so this is not analyzed in this thesis.

The low voltage converters on board have two kinds: the switch DC-DC converters and the Low Drop Out regulator (LDO). Both of them will be discussed in the next chapter. The power supply can be integrated into front-end ASICs or they can be realized out of chip.

The easier way to realize power supply is to buy the commercial shelf products (DC-DC converter or LDO) from the semiconductor companies, such as TI, Linear, ADI and so on. In normal applications this method is enough to get a good power supply. But this kind of way has some drawbacks for HEP applications which have some special features.

The first drawback is the common power supply components which cannot bear the high radiation environment. Of course the high-quality class for the space level has the hard radiation ability. But this kind of components is very expensive and has a long time supply circle. On the other hand, this kind of radiation-aware components sometimes are forbidden by the United States to be used in the non-NATO organization such as China, Russia and so on, even when these components are not used in space or military, such as the accelerator.

When radiation is very high, there is not such kind of the power supply shelf component. For example, the power supply used at LHC, which will run more than 10 years, can accumulate up to 1 *Grad*(Si) dose.

One method to deal with the radiation is to design a special power supply ASIC chip using the radiation-aware technology. This ASIC can use the same or different technology as front-end ASICs.

The second drawback is that the off-chip power supply cannot meet the requirements of excellent performance. The bonding wire may have a few nanofarads of inductance, which can produce one large disturb to the power supply. In these situations, the power supply circuits, designed into the front-end ASIC, are necessary.

A good power supply is one of the determines to get a good performance in front-end ASICs. The aim of power supply is a fine voltage which can supply enough current. The bias voltage and bias current builds on a good performance of power supply.

Nowadays, front-end circuits of nuclear and high energy physics are often the mixed circuit, which not only includes the preamplifier, the shaper, the baseline holder but also includes DCA, ADC, LVDS and so on. The digital circuits are driven by high-frequency clock, which often generates the noise to common lines shared with the analogue circuits. So it needs the power supply having good robust for the disturb caused by digital circuits.

The switch DC-DC and the low drop out voltage regulation (LDO) are normally used in power management. The switch DC-DC has higher effectiveness but low load transient effect, the large ripple. Contrasting this, the LDO has better load transient effect but the effect depends on the output voltage over the input voltage and the ripple is small.

One good choice is the combination of the switch DC-DC and the LDO. The switch DC-DC finishes the first level voltage converter with higher effectiveness, and the LDO finishes the second voltage converter with higher quality. It is shown in Figure 3.5.

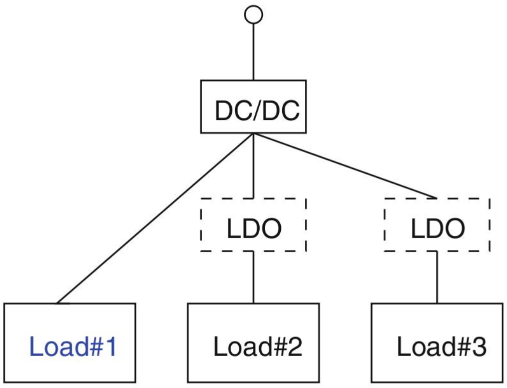


Fig. 3.5: Combination of the switch DC-DC and the LDO

The voltage reference is another important auxiliary component, and it can output the precise voltage which will not sensitive to the input voltage and temperature. The difference between the voltage reference and LOD is that LDO can output the larger current, but the voltage reference has a limited current output with good precise.

The voltage reference is normally realized by bandgap. The purpose of voltage reference is giving the reference to switch DC-DC, LDO, current source or some other circuits, which use the reference getting the precise voltage, current or other important parameters.

3.4 Auxiliary Power Management in Front-End ASICs

As said in previous, the auxiliary power management includes on-chip power and ground distribution grids, bonding wire, pads distribution, the potential of the substrate and nwell connectors and so on.

Normally, front-end circuits include both analogue and digital circuits. Some digital and analogue circuits use different power lines, but they share the same substrate. The noise generated by digital circuits can compromise the performance of analogue circuits [2, 14–19], as shown in Figure 3.6.

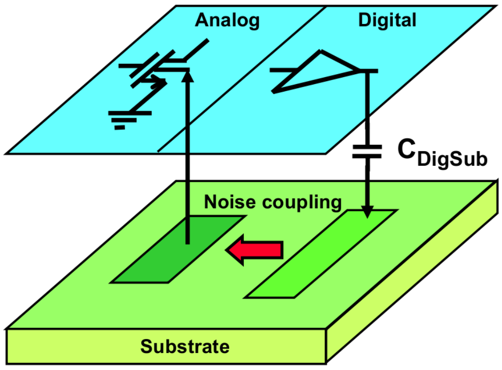


Fig. 3.6: Noise couple through the substrate

The digital circuits inject some noise into the substrate through three methods. The first one is that when digital output states change, some electrons (in the electric field) accelerate and produce the impact ionization effect if the electron speed is enough high. The impact ionization produces the electron-hole pairs and the holes will be absorbed by the substrate.

The second one is that the output of digital will couple the signal into substrate through C_{DigSub}, which is formed by the reverse junctions between the output and substrate.

The third interference pathway is as follows. When switching action happens, the power and ground will charge or discharge electrodes quickly, so there will be fast current going through the pads, which have inductance caused by the bonding wires and the power line. There will be voltage variation including the inductor and resistor voltage drops. The third pathway is the main disturb [16], and it becomes more important in deep sub-micron technology [18].

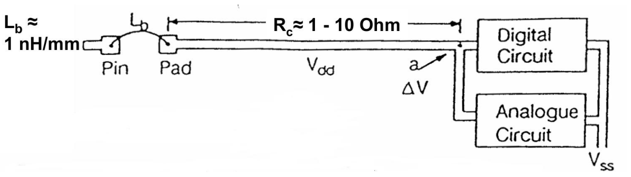


Fig. 3.7: Power supply influenced by the digital circuit

The inductance of bonding wire is about 1 nH/mm. The current going though the wire bonding will generate the bounce voltage [2], as shown in Figure 3.8.

$$\Delta V \approx R_C \cdot I_{dd} + L_b \cdot \frac{\partial I_{dd}}{\partial t} \tag{3.2}$$

where R_C is the resistance of power line, L_b is the bonding inductance, and I_{dd} is the transient current caused by digital switching.

The higher digital clock frequency, the more deteriorating for analogue design. The disturbing voltage is also proportional to the bonding wire inductance as shown in Figure 3.8. Given the bonding wire is about 5 mm length, the bonding inductance is about 5 nH. R_C is normally from 1 Ω to 10 Ω, and here sets it 5 Ω. When the current varies from 10 mA to 20 mA within 1 ns, the $\Delta V = 100$ mV. It is harmful for analogue circuits.

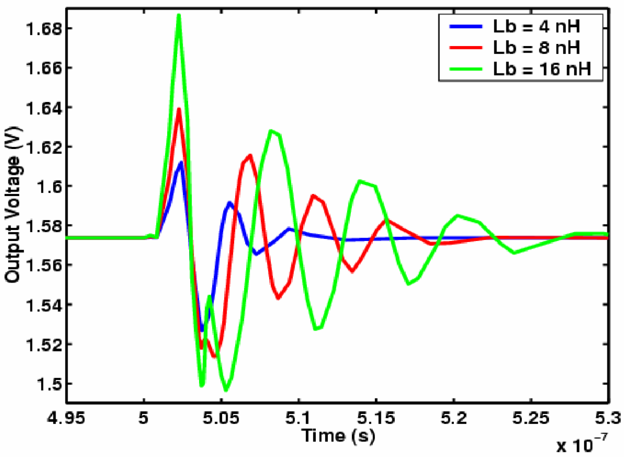


Fig. 3.8: Different bonding wires inductance simulation results

So the shorter bonding wire, the better for the power supply quality. Another method is using the flip-chip package to replace the bonding wire package [20].

From Figure 3.7, it can be seen that the analogue circuit power supply is easily influenced by digital circuits if they share the common power supply. So normally the analogue and digital circuits had better use the separated power supply lines. But in this case, they cost extra pins and area. Sometimes the pins are limited, especially in front-end circuits, which cost more pins by multichannel. In this situation, the digital and analogue circuits have to share one pad. Figure 3.9 shows the comparing between three kinds of pad strategy [2].

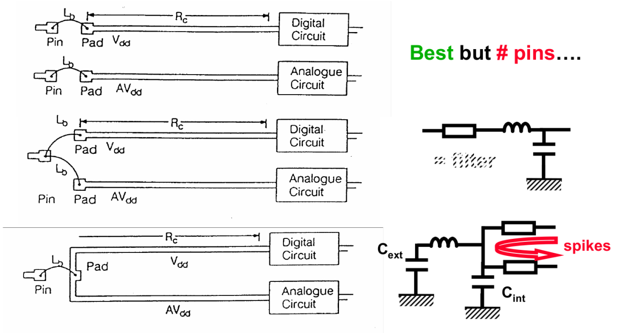


Fig. 3.9: Pads strategy between the analogue and digital circuits

In Figure 3.9, the first pad layout is the best one, which has been said before. The middle one can also be accepted, because there are the resistor, inductor and capacitor to filter the disturb. The last one is the worse one, because it does not make full use of the pins and the external decoupling capacitors.

The above content is about how digital circuits to generate the noise into the substrate. Not only digital circuits, some analogue circuits can also generate the same kind of noise. Such as the fast comparator output signal, which has the same noise production process. Another is the class AB output. When the class AB begins work, the current variation from the quiescent to active state is very large. It can also produce the same noise effect.

From Figure 3.10 a), it can be seen that the voltage bounce in power line pads can directly transfer the noise to the nwell connector. Then it can be transferred to substrate through couple capacitors between the nwell and substrate. The voltage bounce in ground line pads can directly be transferred to substrate, which connects to ground lines.

In order to reduce the substrate influence, the substrate connector and the nwell connector had better own their self-pads, like Figure 3.10 b). Although V_{sub} and the ground will connect off-chip at the end, the voltage bounce caused by the bonding

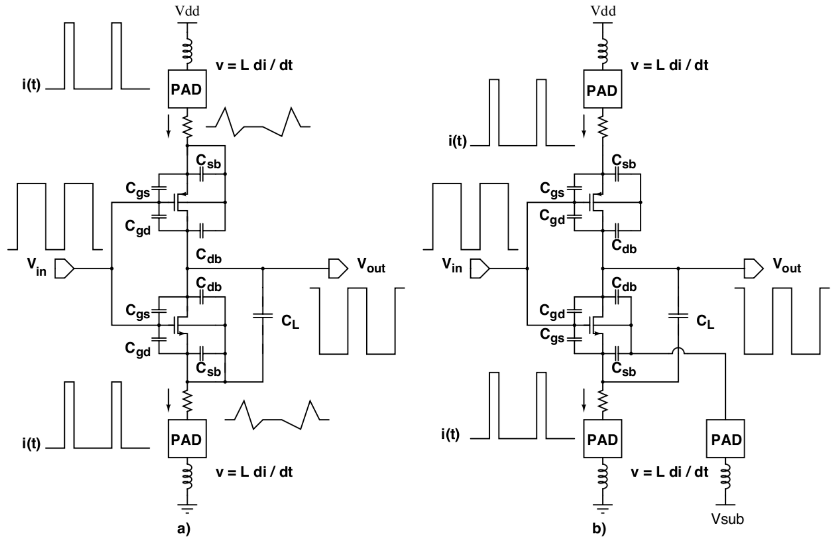


Fig. 3.10: Noise couple caused by the bonding wires

wire and the wire resistance will reduce a lot. Even the substrate and nwell also cost two other pins.

To the sensitive transistors, such as the input transistor of preamplifier, there are also three methods to influence it. The first one is changing the bulk potential which is the fourth terminal of the transistor. That disturbing the bulk potential means to change the threshold voltage, and then transfer the disturb into the drain current. Another is to use the coupling capacitor of junction capacitance. The voltage noise can couple to the nwell, then disturb the sensitive transistor gate or source. The third is the substrate noise transferring to the ground line, which connects to the substrate on-chip or off-chip. On-chip connector have a larger disturb.

Based on the above analysis, for the multichannel front-end circuit design, there are the following notes which must be considered.

The first one is the first stage of preamplifier, which includes the input transistors. It needs separated power supply, ground, the substrate bias and the nwell bias. It is better to have separated pads. It can reduce a lot of digital disturb and the analogue fast signal disturb.

The second one is that the analogue and digital circuits should be separated as much as possible. And it is better to shield the analogue block by the guard rings and the special isolation trenches. Although the whole substrate is at the same potential. Separating the different blocks is also useful for not affecting each other.

The third one is the reasonable arrangement of the power and ground distribution. Reducing IR drop as much as possible. To reduce EMC, the power and ground lines should be alternated arrangement. And there should be enough pads to get a lower bonding wire inductance.

The fourth one is the off-chip decouple capacitors connecting to the power line, which should be as near as possible to power pads. If there is empty space on-chip, this empty space should fill capacitors to decouple the power line for the low noise.

Chapter 4

The Transistor Design of the LDO and Bandgap IP Blocks Using 65 *nm* CMOS Technology

The front-end power supply is one of the key components that affect circuit performance. As the increasing integration degree of the front-end ASIC of high energy physic such as LHC at CERN or BESIII at IHEP, the request of integrating power management with the front-end ASIC increases. Normally power supply of a front-end is voltage supply. There is two types: switching voltage converter and low-dropout voltage supply (LDO).

Regardless of the power supply, an accurate reference voltage is required to obtain a fixed voltage, which is insensitive to input voltage and temperature. This reference voltage is typically realized by a bandgap voltage. The classic bandgap voltage is about $1.25\ V$, which is almost the same as the silicon bandgap voltage.

This chapter and the next chapter will focus on switching DC-DC, LDO and bandgap reference. These three circuits are used to form a power management. The switch DC-DC changes the input high voltage to a lower voltage to complete the first conversion; then the LDO changes the switching DC-DC output to the desired voltage to complete the second conversion. The bandgap output voltage is used as a reference for LDOs and switch DC-DCs. This power management has high power efficiency (because of the switching DC-DC) and high quality power output voltage (because of the LDO).

4.1 The Design of LDO

4.1.1 Introduction

In high energy physics applications, low dropout regulators (LDOs) are widely used in front-end circuits, especially analogue and hybrid ASIC. Comparing with switching conversions, LDOs are simple. Without inductors, LDOs are more compact and therefore smaller and cheaper (consuming less area).

At the same time, LDOs have excellent output ripples performance, so they are more suitable for the front-end of HEP applications. Of course, LDOs have some drawbacks. The main problem is that the LDO efficiency depends on the ratio of the output voltage to the input voltage. Fortunately, this shortcoming can be released by one former switch DC-DC conversion.

Most front-end circuits use CMOS technology which has excellent cost effective. So in this chapter, CMOS technology is used to design LDOs.

In high energy physics applications, analogue and digital circuits are often designed together in the front-end design. A power supply is therefore required, which has very little ripple and fast step response recovery time. And it should avoid digital circuits influencing analogue circuits. To reduce the amount of heat generated, the power supply should have higher efficiency. From the above, LDOs used in HEP applications should have fast recovery time, low dropout, and low noise. It is shown in Figure 4.1.

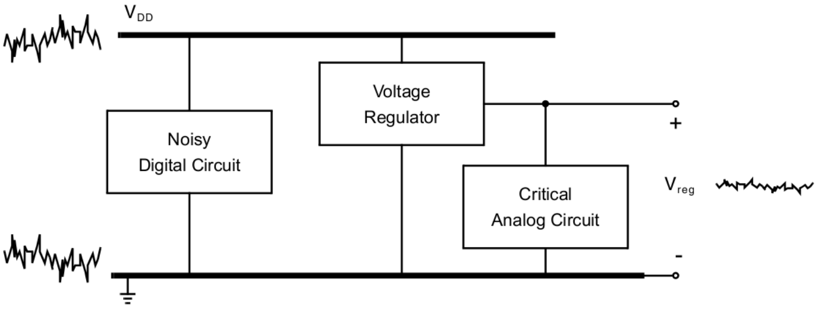


Fig. 4.1: LDO output to the analogue circuit with lower noise

In the traditional LDO architecture, it is difficult to meet these demands. So in this paper, a new LDO topology is introduced. This LDO architecture can be suitable for different output capacitors, regardless capacitance value and the parasitic ESR.

The drop out voltage can be less than 100 mV. When the input is 1.5 V and the output is 1.2 V, the LDO and switching DC-DC efficiency are at the same level. This LDO can provide more than 0.6 A current.

4.1.2 Conventional Linear Regulator Topology

A conventional LDO consists of a passive element called power transistor, an error amplifier, a voltage reference usually produced by one bandgap on the negative terminal of an error amplifier, and a feedback from a voltage divider resistor, as shown in Figure 4.2. In order to obtain a lower drop-out voltage, it uses a PMOS as the power transistor. The divided resistors can provide a series-shunt negative feedback to the error amplifier positive terminal.

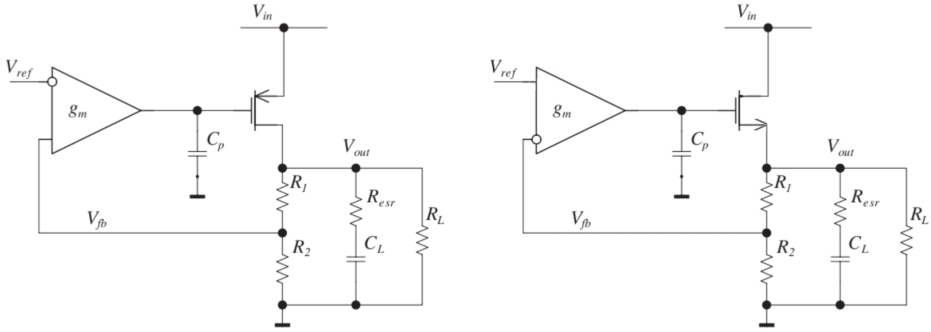


Fig. 4.2: Classical LDO architecture: PMOS (left) and NMOS (right) power transistors

There are two types of power transistors: NMOS transistors and PMOS transistors. They have different pros and cons. The NMOS power transistor has lower output impedance $1/g_m$ and has fast response to load current variation. But the drawback is the higher dropout $(V_{ds}+V_{gs})$. When using one PMOS as the power transistor, the drop out will be low to V_{ds} and thus improve high power efficiency. But the output impedance will be high up to R_{ds} which is larger than $1/g_m$, therefore, its load response capability is very poor. It can refer to Table 4.1. So a new LDO architecture is needed, which can improve the contradiction structure.

The maximum current of the LDO depends on the size of the power transistor. When the maximum current is several hundreds of mA, the width of the power transistor may reach several millimeters. In this case, the gate parasitic capacitance will be large enough that the output of error amplifier has a low frequency pole.

Table 4.1: Different power transistors

Parameter	NMOS	PMOS
$\mathrm{V}_{\text{dropout}}$	$\mathrm{V}_{\text{SD(sat)}} + \mathrm{V}_{\text{GS}}$	$\mathrm{V}_{\text{SD(sat)}}$
$I_o(\max)$	Low	High
R_o	$1/g_{m(\text{MOS})}$	r_{ds}

At the same time, the output filter capacitors are often large value, so it will produce another low frequency pole at the output terminal. In this case, the LDO system will have two low frequency poles, and thus it will be unstable. In order to keep LDO systems stable, conventional LDOs have two compensation circuits: an external compensate circuit and an internal compensate circuit [21].

When using an external compensation circuit, the main pole is at the system output. The error amplifier output will be the second pole. So the output terminal requires a large value capacitor and it must be larger than a certain value. Because of the large capacitance value, output filter capacitors are often placed outside the chip.

Compensate capacitors should have a high ESR, which can play a low zero point. A low ESR (such as the high-density ceramic capacitors) will introduce a high zero point, if the high zero point is large than the unit-gain frequency, the zero point will not play a role.

ESR is a parasitic resistance, and the different values depend on the process. It is therefore difficult to control the ESR to a specific value. The conventional LDO with the external compensation capacitance is not suitable for high energy physic front-end which requires small area, especially for the multichannel detectors.

When using internal compensation, the main pole will be at the output of the error amplifier, and the second pole will be at the output of the system. So the compensation capacitor should not exceed a certain certain value. In this case, when the load current changes, the output voltage needs more time to recover than the external compensation. See the Table 4.2 for their advantages and disadvantages.

When analogue and digital designs are combined, the output load current often changes a lot due to the switching action of the comparators and digital circuits. So the internal compensation is not suitable for HEP front-end applications.

External compensation has better performance, but it requires a large output capacitor with some ESR, so it is also not suitable for the ASIC design. In order to

Table 4.2: External and internal compensations table

	External compensation	**Internal compensation**
Dominate Pole	Output pole p_o	Error-amplifier pole p_a
C_o **Limit**	Greater than $C_{specified}$	Less than $C_{specified}$
Load-Dump Response	Better,lower ΔV_{OUT}	Worse, larger ΔV_{OUT}
PSR(or A_{IN}^{-1}**)**	Better	Worse
Integration	Off-chip C_o	On-chip C_o
Application	**Higher power, larger load dumps**	**Lower power smaller load dumps**

get a better LDO for front-end circuits, a new LDO is essential, which does not care the output capacitance load and has a low dropout voltage.

4.1.3 Circuit Design Description

Most classic LDO devices use external compensate, because of their excellent performance. With the development of sub micron CMOS technologies used in high-energy physics front ends, LDOs are often used in large numbers. The classical external compensate LDO cannot be used, because every such LDO requires an external capacitor and the dedicated output pins. Typically, there are not many areas and pins in the front-end CMOS ASIC. Using this structural approach, an any-load stable LDO is created [22], as shown in Figure 4.3.

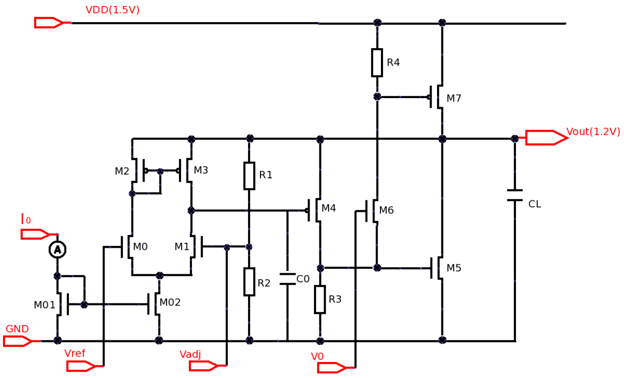


Fig. 4.3: A new LDO architecture

In this circuit, the four transistors on the left (M_0, M_1, M_2, M_3) and the current source I_0 consist of an error amplifier, and the M_4, M_5 consist of a super source follower. Because there is a negative feedback in the super source follower [23], the output impedance becomes smaller from $1/g_{m4}$ to $1/(g_{m4} \cdot g_{m5} \cdot \mathrm{R}_{o4})$.

The error amplifier and super source follower can be used to build a two stages system, which can be unconditionally stable with any capacitor load when limiting the first stage voltage gain. The method to limit the first stage gain is reducing the impedance of the M_4 gate. So M_0/M_1 and M_2/M_3 length should be reduced [24, 25].

By the same principle, in the super source follower, in order to get the unconditional stability, the voltage gain of M_4 should be controlled, which is about $g_{m4} \cdot \mathrm{R}_{p5}$. The R_{p5} is the equivalent resistance at the gate of M_5. Lower R_{p5} can be gotten through shortening of the M_4 channel length. The function of C_0 is not compensation. It is used for steadying the gate of M_4 in order to make M_4 as one common gate transistor when M_4 and M_5 form the negative feedback.

Another difference with the conventional LDO is the power supply of error amplifier. In this architecture, the error amplifier is powered by $\mathrm{V}_{\mathrm{out}}$, not by the VDD, so VDD will have less influence on the error amplifier. It makes the system have better PSRR.

According to the above analysis, the appropriate transistor size is obtained as shown in Table 4.3 (using TSMC 65 *nm* technology).

Table 4.3: Transistor sizes for the circuit of the new LDO (W/L unit: μm; R unit: Ω)

$M0 = 20/1$	$M1 = 20/1$	$M2 = 15/1$	$M3 = 15/1$
$M4 = 800/0.6$	$M5 = 12/0.6$	$M6 = 144/0.6$	$M7 = 57600/0.2$
$M01 = 20/1$	$M02 = 20/1$	$R1 = 10\ k$	$R2 = 10\ k$
$R3 = 150$	$R4 = 400$		

4.1.4 Radiation-Aware Design

This thesis designs three kinds of LDOs. One uses standard NMOS transistors, one uses H shape NMOS transistors, and the last one uses ELT NMOS transistors. In the future, when there is an opportunity to perform radiation testing, the three LDOs can be compared to each other, and the anti-radiation capability of the H shape and ELT transistor can be obtained. The radiation effects on the CMOS technology LDO are TID and SET.

LDOs are implemented by the TSMC 65 *nm* CMOS technology, and according to Chapter 2, the technologies below 90 *nm* are less sensitive to TID radiation. The advanced technology can reduce the threshold voltage shifts caused by TID. Another radiation-Aware method is to place the guardrings around most transistors in order to reduce the leakage current caused by TID.

Besides the methods for TID, when doing the layout, as much as the substrate connectors and nwell connectors are placed, in order to reduce the parasitic resistance of the parasitic triodes. The purpose is reduce SET chance.

As described in Chapter 1, Menouni and his colleagues [4] at CERN developed a TID radiation transistor model for the standard transistors. So the standard transistor LDO can be simulated using the radiation model. This model includes 200 $Mrad$(Si) and 500 $Mrad$(Si) transistors. Using this model transistor to simulate, the result can show the LDO TID radiation effect under 200 $Mrad$(Si) and 500 $Mrad$(Si). Here simulates the load variation, which is one of the key parameters. The result is shown in Figure 4.4.

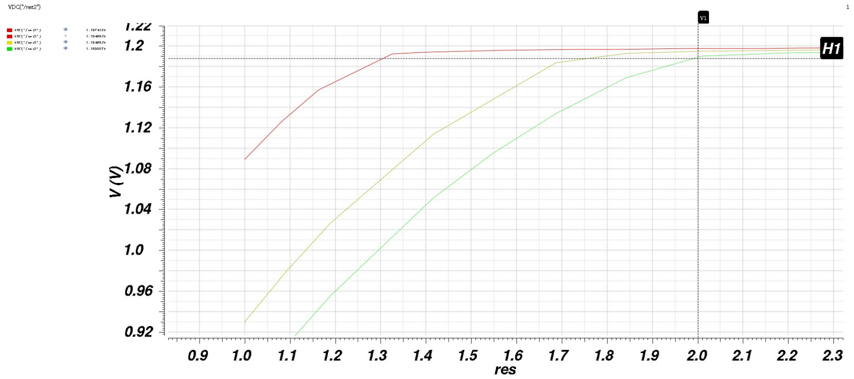


Fig. 4.4: LDO radiation simulation comparison

The figure shows that before the radiation the maximum simulation current is about 800 *mA* (when testing the actual LDO, the value may be reduced because of parasitic parameters). After 200 $Mrad$(Si), the simulation maximum current will drop to 690 *mA*, and after 500 $Mrad$(Si), the simulation maximum current will drop to 600 *mA*. 200 $Mrad$(Si) is already a large number, which can only happen in very high energy experiments, such as the LHC. In the BESIII experiment or normal space experiments, the TID is less than 300 $krad$(Si). So the LDO can withstand the TID radiation effects in BESIII experiment. According to the TID production mechanism, the H shape LDO and ELT LDO should have better anti-radiation ability.

4.1.5 Simulation and Layout

The new architecture LDO is realized by TSMC 65 *nm* technology. The input is 1.5 *V* and output is 1.2 *V*. The output can be adjusted from 1.1 *V* to 1.3 *V* by the parallel resistors with the feedback divided resistors.

One of the other important parameters is the load transient response, especially in the analogue and digital hybrid circuits. Since the digital circuits can cause the load current variation a lot, this load variation can affect the power line voltage and then affect the analogue power supply. Through simulation, when the load current changes between 0 to 0.6 *A*, the recovery time is less than 200 *ns*, as shown in Figure 4.5.

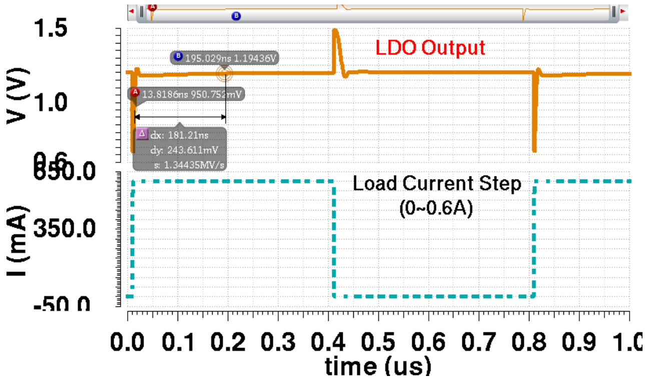


Fig. 4.5: Load transient response

Dropout voltage is the minimum voltage different between the input and output voltage when LDO work normally, so it determines the LDO highest efficiency. The dropout voltage has a relationship with the output current. The smaller load current, the lower dropout voltage. So when an LDO outputs the larger current, there will be the fewer dropout.

Using one PMOS transistor as the power device, the drop out voltage can be very low. Through simulation, at the output current (0.6 *A*), the dropout voltage can reach about 100 *mV*. It is slightly larger than the light load dropout voltage (36 *mV*), as shown in Figure 4.6, but it is lower than most commercial LDOs of the same level.

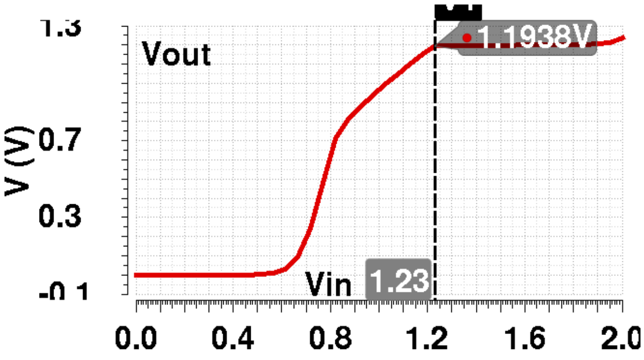


Fig. 4.6: Drop out voltage with output 0 A

Another important parameter is load regulation. This parameter reflects the output voltage variation when the load current changes. A good LDO should have a small load regulation.

Normally there are three regulation comparing standards, different companies choose different standards. One is to compare the output voltage variation when given the solid load current variation, the unit is volt; another is comparing the ratio of the output changed voltage to the original output voltage for a given solid load current change, the unit is percentage; the third is to compare the ratio between the output voltage variation and the load current variation, the unit is the Ω.

Here uses the third standard, and the unit is V/A or Ω. Through simulation, the load regulation is about 1.47 mV/A. That means when the output current changes 1 A, the output voltage changes 1.47 mV. It is shown in Figure 4.7.

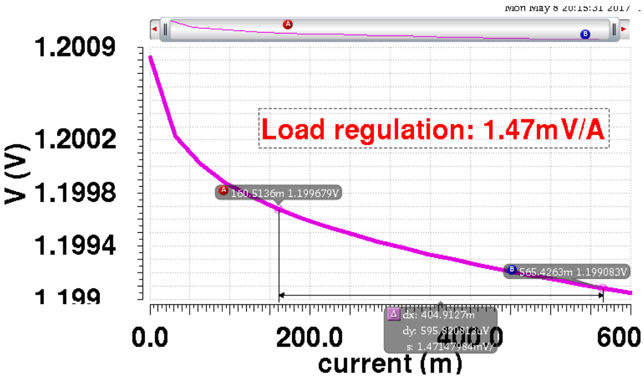


Fig. 4.7: Load regulation of the LDO

The input voltage can also affects the output voltage, and this parameter is line regulation. The power supply of LDO usually comes from other power converters. The power lines have some resistance. Therefore, the power supply may be unstable.

The LDO requires a good line regulation to reject the power supply variation. The line regulation is shown in Figure 4.8. From this figure, it can be seen that the LDO regulation range is from 1.28 V to 1.79 V. When the power supply changes in this range, the output voltage changes from 1.1971 V to 1.2007 V. The LDO line regulation is about 7 mV/V or 0.7%.

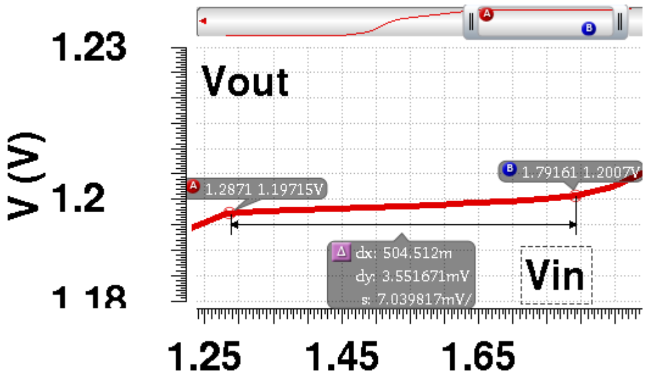


Fig. 4.8: LDO line regulation

Power supplies usually come from switch DC-DC converters and is connected to other digital circuits. The switching DC-DC output ripples are often large and digital circuits always produce some noise back to the power supply. So it needs LDOs to filter this ripple or noise. If LDOs have a bad line regulation, these power supply noise will shift to the LDO output and therefore influence output voltage quality. Based on this reason, LDOs should have better PSRR. A good PSRR should be larger than 40 dB. The LDO PSRR is shown in Figure 4.9. From the figure, at low frequency, the PSRR is about 43 dB, which is good. But at the middle and high frequency, the PSRR is not good, which need to be improved on the next step.

Layout of LDOs

Power management requires three LDOs to provide the front-end ASIC. One is used for the pre-amplifier, one is used for the other analog circuit, and the last one is used for the digital circuit. Three supplies can reduce the noise crosstalk between different circuit parts. In order to compare the different LDOs anti-radiation capability, this

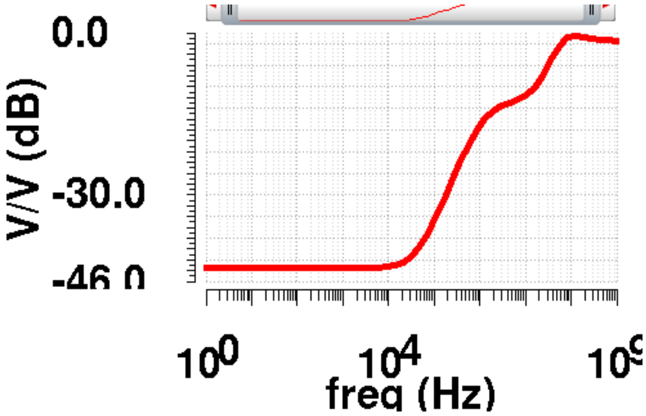


Fig. 4.9: PSRR of the LDO

thesis uses three kinds of NMOS for the LDOs: the standard transistor, the H shape transistor and the ELT transistor. When doing the radiation test in the future, the different LDO appearance can be gotten in different radiation environments. The H shape and ELT transistors are shown in in Figure 4.10.

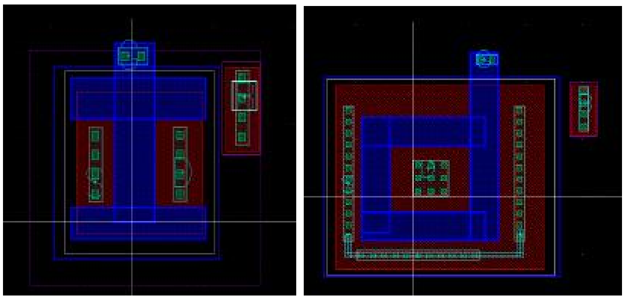


Fig. 4.10: The H shape (left) and ELT (right) transistors

The three LDO layouts are almost identical except for the NMOS transistors. So here only the H-shaped LDO layout is given as an example, as shown in Figure 4.11.

This layout shows that approximately half of the silicon area is occupied by the capacitor, and power PMOS transistor occupies most of the left silicon area. A key point of a large PMOS is to output a larger current. Another key point is to reduce parasitic resistance. So the large power PMOS consists of many small width parallel transistors.

Large capacitance, which need large area, is independent of stability, but it can reduce fluctuations in transient load voltage. So the output capacitance occupies much area. In order to convenient layout, three LDOs are located at three different corners, where can use more pads for the large current.

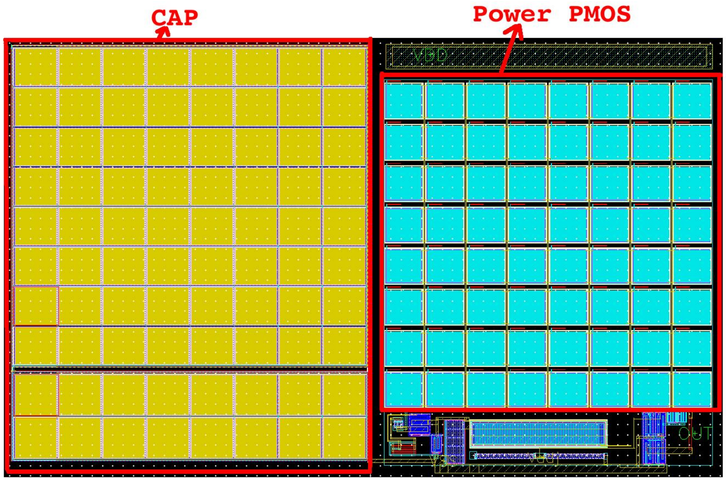


Fig. 4.11: The H shape LDO layout

4.1.6 Conclusions

To compare this work to other LDO designs, FOM (figure of merit) is required, with a focus on recovery time, power current density, and dropout voltage. These parameters are key parameters for high energy physics applications.

$$FOM = K \cdot \frac{V_{outpp}}{I_{outpp}} \cdot I_q \cdot V_{do} \tag{4.1}$$

where K is the power current density, $K = Area/I_{maxcurrent}$. V_{outpp} is the output voltage peak to peak voltage; I_{outpp} is the output maximum test current variation; I_q means the quiescent current; $V_d o$ is the drop out voltage.

When using this FOM to compare LDOs, the less FOM value, the better suitable for front-end circuits. The performance of this work and other designs are shown in Table 4.4.

As can be seen from the above table, this work performance is suitable for high energy physics design. The main purpose of this work is to examine this new architecture in 65 *nm* technology. It will work together with the switching DC-DC for the power management. At the same time, this work can do the technical reserves for the future.

Table 4.4: Parameters comparing of different LDOs

Parameter	**Unit**	[26]	[27]	[28]	[29]	this work
Year		2013	2015	2016	2016	2017
Technology	μm	0.11	0.18	0.5	0.18	0.065
Vin	V	$1.8-3.8$	$1.4-1.8$	$2.3-5.5$	$1-1.4$	$1.4-2$
Vout	V	1.2	1.2	$1.2-5.4$	0.9	$1.1-1.3$
Iout(MAX)	mA	200	100	150	0.5	600
Iq	mA	41.5	141	40	10.3	3000
Vdropout	mA	200	100	150	0.5	90
Undershoot	mV	385	110	96	64	500
Overshoot	mV	200	85	120	64	300
Voutpp	mV	585	195	216	128	800
Ioutpp	mA	199.5	99.99	150	0.25	600
Set time	μs	0.65	30	3	3	0.2
Area	mm^2	0.11	0.07	0.279	0.012	0.04
FOM	μs	13.4	38.5	11.4	126	12.5

4.2 The Design of Bandgap

4.2.1 Introduction

Bandgap is a voltage reference commonly used for ADC, LDO, switching DC-DC conversion and so on. The bandgap reference will provide a reference voltage, which is insensitive to the power supply and temperature. The normal bandgap provides 1.25 V voltage, which is almost the same as the silicon bandgap voltage. This is the origin of its name.

Because temperature variation can influence most process parameters. If circuits are not sensitive to temperature, they will not be sensitive to process parameters. So the temperature insensitive will be an importance parameter. The bandgap voltage is often obtained by the two voltage combination: one with a negative temperature coefficients and the other with a positive temperature coefficients.

Negative temperature coefficients (Negative-TC) are usually derived from diodes or diode like components (such as the diode-connected triode, the DTMOS and the sub-threshold CMOS), which all have an exponential function between the voltage and the current. The diode current is:

$$I_{diode} = I_s \cdot exp(V_{diode}/V_T) \tag{4.2}$$

where I_s is the saturation current, V_T is the thermal voltage kT/q. The V_{diode} is:

$$V_{diode} = V_T \cdot \ln \frac{I_{diode}}{I_s} \tag{4.3}$$

According to the book [30], the derivation of the V_{diode} temperature is

$$\frac{\partial V_{diode}}{\partial T} = \frac{V_{diode} - AV_T - B}{T} \tag{4.4}$$

where A and B are constant. As can be seen from the equation, the V_{diode} temperature coefficient (TC) is related to V_{diode} and absolute temperature. When the positive TC is constant, the entire TC will have some deviation. In room temperature, the Negative TC is about -1.5 *mV* to 2 *mV*/degree, which depends on the process.

Positive TC can be obtained by the difference of the two diodes voltages, which have different current. It can be seen in Figure 4.12. One branch current is nI_o which is n times the other branch current I_o. The difference of two diode forward voltages is:

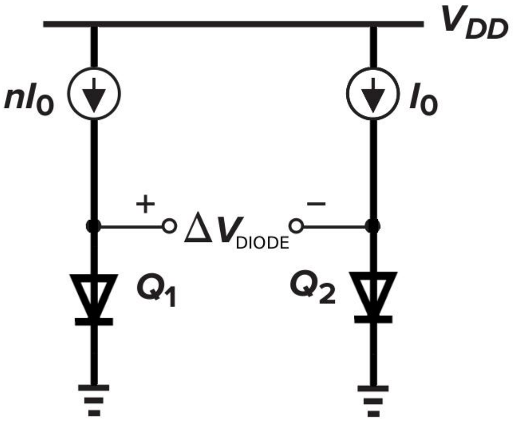


Fig. 4.12: Positive TC architecture

$$\Delta V_{DIODE} = V_{diode1} - V_{diode2} = V_T \cdot \ln \frac{I_{diode1}}{I_s} - \frac{I_{diode2}}{I_s} = V_T \cdot \ln n \tag{4.5}$$

When the Positive TC circuit and the Negative TC circuit work together, it can obtain almost independent temperature circuit, as shown in Figure 4.13.

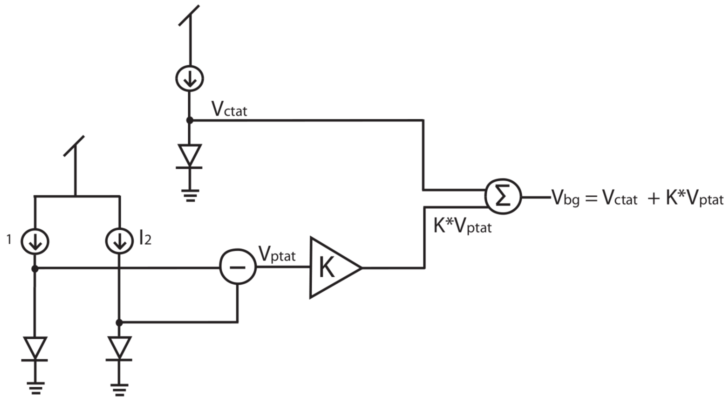


Fig. 4.13: Bandgap principle diagram

where PTAT is proportional to absolute temperature and CTAT is inversely proportional to absolute temperature. When setting the parameters to make the two complementary TC cancel each other, the output voltage will not sensitive to temperature.

To realize PTAT, the key feature is the voltage and current of device with an exponential relationship. Besides the diodes, the diode-connected triodes, the sub-threshold transistors and the dynamic threshold transistors (DTMOST) also have such feature and they can be used in the bandgap. CMOS technology can realize the diode on-chip, but diodes often have one 1.00-1.05 uncertain V_T coefficient. The diode-connected triodes do not have such a V_T coefficient. Therefore, CMOS technology usually uses the diode-connected triode for the bandgap.

Some kinds of special triodes can be realized by CMOS technology, as shown in Figure 4.14. These triodes are parasitic triodes including lateral PNP, lateral NPN and vertical PNP.

Among the three kinds of parasitic triodes, the gate of vertical PNP triode is thinner than the lateral triode. So the vertical transistor has the larger current gain coefficient β, and it has more gain than the other two lateral triodes. It is important to note that when using the parasitic triodes, it cannot use guard rings. Since guard rings can reduce β, which makes the gate absorb more electrons from the emitter.

Based on this reason, vertical PNP transistors are more commonly used in CMOS technology. However, when using this vertical PNP, the collector terminal must be

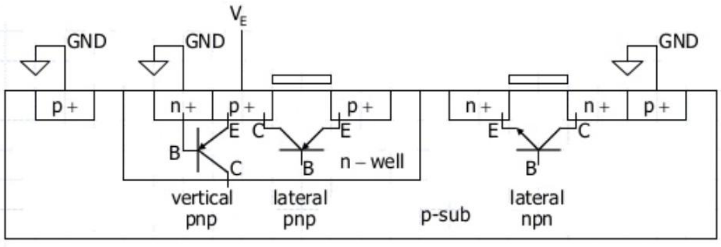


Fig. 4.14: Special triode using CMOS technology

restricted to the ground, which limits its application, however, it is suitable for the classical bandgap. It is shown in Figure 4.15.

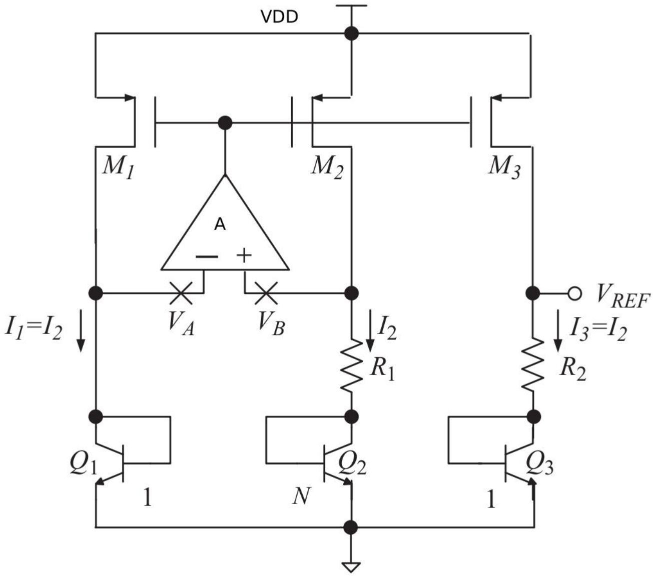


Fig. 4.15: Classical bandgap technology

In this bandgap, the vertical PNP transistors are implemented by PMOS, and the amplifier forms two feedbacks. One is positive feedback, and the other is negative feedback. The amplifier A, M_2, R_1 and Q_2 constitute negative feedback: when V_B increase, the output of A rises, which reduces the M_2 current I_2. The reduced I_2, the resistance of R_1 and Q_2 decrease V_B. The negative feedback coefficient is:

$$\beta_N = g_{m2} \cdot (R_1 + \frac{1}{g_{m,Q2}}) \tag{4.6}$$

Using the similar method, the positive feedback consists of the amplifier A, M_1 and Q_1. The positive feedback coefficient is:

$$\beta_P = g_{m1} \cdot \frac{1}{g_{m,Q1}} \tag{4.7}$$

In order to keep the system stable, the negative feedback coefficient must be larger than the positive feedback coefficient, in order to make the whole system negative feedback. It is better to choose $\beta_N \approx 2 \cdot \beta_P$, in order to have a fine transient response even there is a large capacitance on the output [30]. Due to the system negative feedback, V_A and V_B will maintain the same potential.

When V_A equals to V_B, and the M_1, M_2 have the same size, the current I_1 and I_2 are the same. Q_2 is N times Q_1, so the Q_2 current is about $1/N\ Q_1$. There will be PTAT between the Q_1 gate voltage and the Q_2 gate voltage V_A. V_B is the same as V_A, so the R_1 voltage will be PTAT. And then the R_1 current I_2 is also PTAT.

$$I_2 = V_T \cdot \frac{\ln N}{R_1} \tag{4.8}$$

When M_3 has the same parameters as M_2, M_3 current I_3 will be the same as I_2. When PTAT I_3 flowing through R_2, there will be the PATA voltage produced by R_2. Adjust to suitable ratio, the PATA R_2 voltage is added to the CATA voltage V_{Q3} to get the temperature nonsensitive output V_{REF}.

$$V_{REF} = I_2 \cdot R_2 + V_{BE3} = \frac{R_2 \cdot \ln N}{R_1} \cdot V_T + V_{BE3} \tag{4.9}$$

In this equation, $V_T = k_B \cdot T/q$ has the positive TC, and V_{BE3} has the negative TC. In room temperature the V_T temperature coefficient is:

$$\frac{dV_T}{dT} = \frac{d(k_B \cdot T/q)}{dT} = \frac{k_B}{q} = \frac{1.38 \cdot 10^{-23}\ J/K}{1.6 \cdot 10^{-19}\ C} = 0.086\ mV/K \tag{4.10}$$

The positive TC depends on the current and process, and its range is from -1.5 mV to -2 mV. So $\ln N \cdot R_2/R_1$ range is from 16 to 22. The larger ratio R_2/R_1 requires less N, vice versa. The V_T is 25.9 mV in room temperature, so the PATA voltage range is from 416 mV to 570 mV. When this voltage is added to a diode voltage of 0.7 V, the whole voltage is approximately 1.2 V. The final output is about 1.2 V, depending on the current and the process.

In most bandgap circuits, there is one "degenerate" bias point. Before the circuits works, there is a stable state, in which no current flows in the circuits. For example, in a classical architecture, when the output of amplifier is at high potential, such as VDD, and the V_A, V_B are at a low potential, such as ground. In this case, the circuit can remain stable and no current flows through the bandgap, and the output voltage is zero. So normally, the bandgap requires a start-up circuit, which will work at the beginning and will no longer work when the bandgap is working normally.

4.2.2 Architecture of Bandgap

The classical bandgap output is normally 1.2 V and requires a power supply larger than 1.2 V. In deep micro process, power supply may just be 1.2 V or less, and output voltage is even less than 1 V. In this situation, the classic architecture will not be suitable.

Based on the classic architecture, this thesis provides two improved architectures, which can be used in deep micro process. They are implemented by the TSMC 65 *nm* technology with the power supply of 1.2 V, and the bandgap output voltage is 600 mV. One architecture uses OPA, and the other one does not use OPA.

Low Voltage Bandgap with the OPA

The classic bandgap complements the PTAT and CTAT voltage to obtain a temperature nonsensitive voltage. There is another way to get the bandgap. The first step is to get the PTAT and CTAT current, and then sets the parameters to make them complementary. In this case, the temperature nonsensitive current can be changed into a temperature nonsensitive voltage through one resistor.

Banba designed such a low power bandgap, as shown in Figure 4.16 [31]. This chapter designs a bandgap based on the Banba architecture.

In this architecture, R_2 and R_3 are set the same. V_A, which is the forward voltage of the diode-connected triode, is a PATA voltage, so the R_2 current I_{1a} is a PATA voltage. When the operation amplifier (OPA) has enough gain, V_A and V_B will be almost the same, and then the current flowing R_2, R_3 will be almost the same.

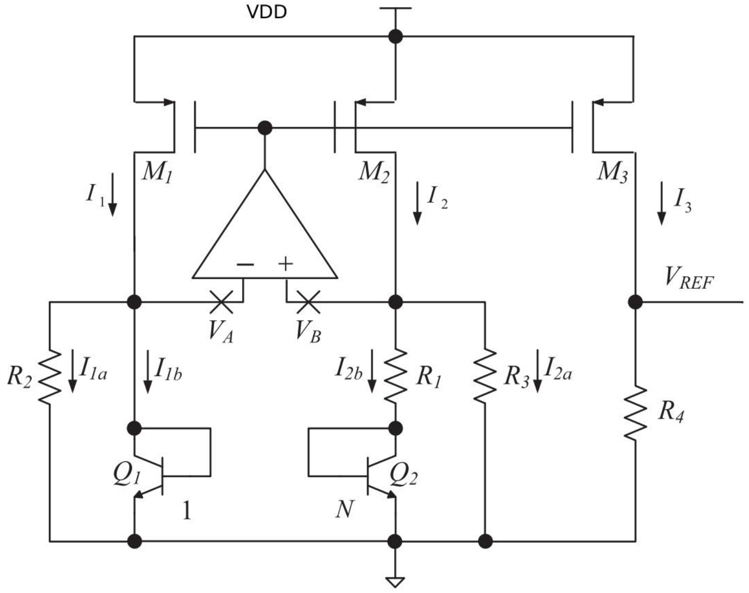


Fig. 4.16: Banba low power supply bandgap architecture

$$I_{2a} = \frac{V_B}{R_3} = \frac{V_A}{R_3} = \frac{V_{BE1}}{R_3} \tag{4.11}$$

When M_1 and M_2 have the same size, the current I_1 and I_2 will be the same. I_{1a} equals to I_{2a}, so I_{1b} equals to I_{2b}.

$$I_{2b} = \frac{V_B - V_{BE2}}{R_1} = \frac{V_{BE1} - V_{BE2}}{R_1} = \frac{\Delta V_{BE1,2}}{R_1} \tag{4.12}$$

$\Delta_{BE1,2}$ is the difference of the two parasitic triodes base-emitter voltage.

$$\Delta_{BE1,2} = V_{BE1} - V_{BE2} = V_T \ln(N) \tag{4.13}$$

N is the emitter area ratio between Q_1 and Q_2. When M_3 has the same dimension as M_1,M_2, M_3 current I_3 is the same as I_2 and I_1. The final output voltage is:

$$\begin{aligned} V_{REF} = I_3 \cdot R_4 = (I_{2a} + I_{2b}) \cdot R_4 = \left(\frac{V_{BE1}}{R_3} + \frac{\Delta V_{BE1,2}}{R_1} \right) \cdot R_4 \\ = \frac{R_4}{R_3} \cdot \left(V_{BE1} + \frac{R_3}{R_1} \cdot \ln(N) \cdot V_T \right) \end{aligned} \tag{4.14}$$

The bracket equation includes the CTAT V_{BE1} and the PTAT V_T, which is like the equation of classical bandgap. To compensate the negative TC of V_{BE1}, the coefficient of PATA V_T is from 16 to 22, and the bracket voltage is about 1.2 V which is nonsensitive to temperature. The output voltage will be divided by R_4 and R_3. When R_4 is small, there will be the small output voltage which is lower than 1 V.

The parasite triode has one drawback, which is sensitive to the radiation especially the lower radiation rate TID. The TID radiation will damage the triode and form absorb centers which recombine the electrons from the emitters. So it can reduce the current gain coefficient β. When used in high energy physics environment with high radiation, it should avoid using the triode.

There are two other methods to design the radiation aware circuits. One is using the sub-threshold CMOS to replace the parasitic triode, the other is using the dynamic threshold CMOS to replace the triode. Because both kinds of transistors have the similar exponential relationship between the current and voltage. They can be used to generate the PTAT voltage and current.

Comparing with NMOS, PMOS has a stronger anti-radiation ability. In this design, the sub-threshold PMOS is used in the bandgap. The transistor level bandgap is shown in Figure 4.17.

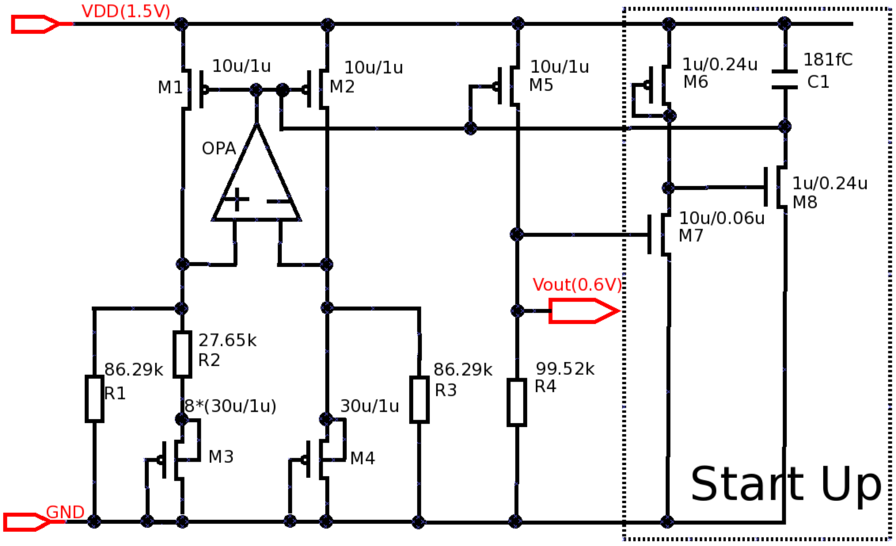


Fig. 4.17: Transistor level bandgap schematic

This bandgap power supply is 1.5 V and the output voltage is 0.6 V. It uses the OPA to make R_1 and R_3 the same voltage. M_3 and M_4 are diode-connected PMOS, and the ratio of them is 8. R_2 changes the PATA voltage into the PATC current. Together with the R_1 CTAP current, there will be the temperature nonsensitive current flowing through R_4, so the temperature nonsensitive voltage can be produced.

When the OPA output is the VDD and at the same time the R_1 and R_3 voltage is zero, the bandgap will not work. The right dotted line includes the start-up circuit, which is used for avoiding this situation. The start-up course is as follows. At the beginning, when the gate of M_5 is at high potential, there is not the current flowing R_4, so the output voltage is zero. M_7 is switched off, and no current flows through M_6. So the drawn of M_7 is high. When the M_7 high drawn voltage switches on M_8, which will absorb the current of the M_5 gate, and make the gate potential decrease. The bandgap begin works. After the bandgap normally outputs 0.6 V, M_7 will be switched on, which makes M_8 switch off, so the start-up circuit will not influence the gate of M_5.

Because of the finite output impedance, the output voltage will be easily influenced. The output PMOS gate drain capacitance will couple the disturb onto the OPA output. To reduce this disturb, the best method is to enlarge the OPA open-loop bandwidth, which makes the OPA reflect more quickly. But this method will consume more power. As a reference circuit, the bandgap should consume small power. Here uses a compromise method, which uses a capacitor C_1 to make the gate voltage stable.

This capacitor C_1 and the parasite gate drain capacitance C_{stray} will divide the disturb voltage V_{dis}. So the gate voltage will change $V_{dis} \cdot C_{stray}/(C_{stray} + C_1)$. So it needs C_1 greatly larger than the parasite gate drain capacitance. But the large C_1 has two drawbacks. One is the large capacitor making the OPA unstable, and the other is the C_1 slowing down the OPA transient response. So choosing C_1 is a trade off. Here chooses C_1 of 181 fC for compromise.

The OPA should have large bandwidth and high gain, so the cascode technology is used. The OPA schematic is shown in Figure 4.18.

To reduce the input error voltage, OPA needs a large open-loop gain at least 60 dB. Because the TSMC 65 nm technology just has 10-20 times intrinsic gain, the OPA uses an folded cascode architecture. Considering the output capacitance, the OPA use a Miller capacitor to stabilize the OPA.

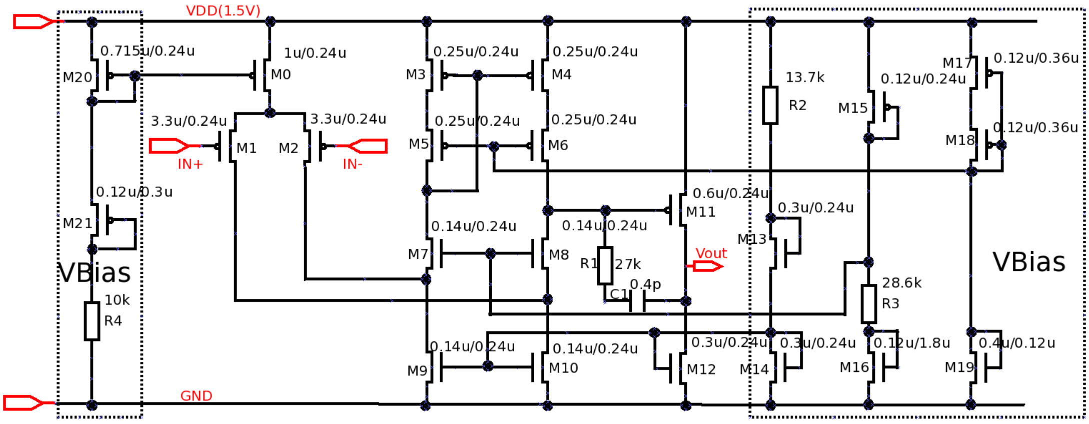


Fig. 4.18: Transistor level OPA schematic used in the bandgap

The dotted black line is the bias voltage circuit. There is one point should be noticed. The power supply is 1.5 V, but the core transistors maximum voltage is 1.32 V. When designing the core circuits and the bias voltage, the transistors voltage should not exceed the maximum value of 1.32 V.

Low Voltage Bandgap without OPA

There is another kind of bandgap, which does not use the OPA. So it will consume very few current, therefore it is more suitable for the low power situation. From the last section Figure 4.16, it can be seen that the function of OPA makes V_A equal to V_B. If there is another circuit to make the two voltages at the same, the PATA current can be also generated. The book [30] gives another circuit to get the PTAT current, as shown in Figure 4.19.

This circuit uses a current mirror to make up the negative feedback, in order to maintain the X,Y point at the same potential. Based on this PATA current, the bandgap can be set up easily. But one of the drawbacks in this circuit is the small feedback coefficient, which can make V_X and V_Y have some difference. To get a larger feedback coefficient, a new architecture is used in Figure 4.20 [32].

This bandgap realizes the reference voltage without OPA. The accuracy of the bandgap reference comes from the matching of the transistor pairs, so these transistors will have large length and therefore have a large area. And all matching transistor should work in the saturation range, which have a better matching effect.

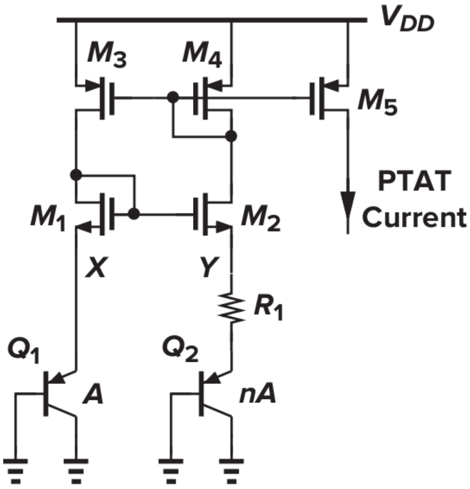


Fig. 4.19: Alternative method of the PTAT current without OPA

In order to improve the loop gain, which can reduce PSRR, this circuit uses the cross-coupling method.

Because no OPA used in the circuit, the power supply voltage can be lower. All current, used to make V_a the same, comes from the diodes bias current, so the current efficiency can reach the maximum. It can realize ultra low power. However, this circuit has some drawbacks. The output voltage depends on the diode negative TC and the output cannot be changed easily. The diodes are not suitable for radiation surroundings.

This chapter combines the advantages of the Banba architecture and the no OPA architecture, then offers a new kind of bandgap, which is shown in Figure 4.21. This new bandgap uses sub-threshold PMOS to get the negative TC, which has the exponential relationship between the current and the voltage.

Like the OPA architecture bandgap, this new architecture also have two kinds of feedback: the positive feedback and the negative feedback. The negative and positive feedback coefficients are:

$$\beta_N \approx g_{m6} \cdot \left(\frac{1}{g_{m8}} + R_3 + R_x \right) \frac{g_{m3} \cdot (R_3 + R_x)}{1 + g_{m3} \cdot (R_3 + R_x)} \tag{4.15}$$

$$\beta_P \approx g_{m7} \cdot \left(\frac{1}{g_{m7}} + R_x \right) \frac{g_{m4} \cdot R_x}{1 + g_{m3} \cdot R_x} \tag{4.16}$$

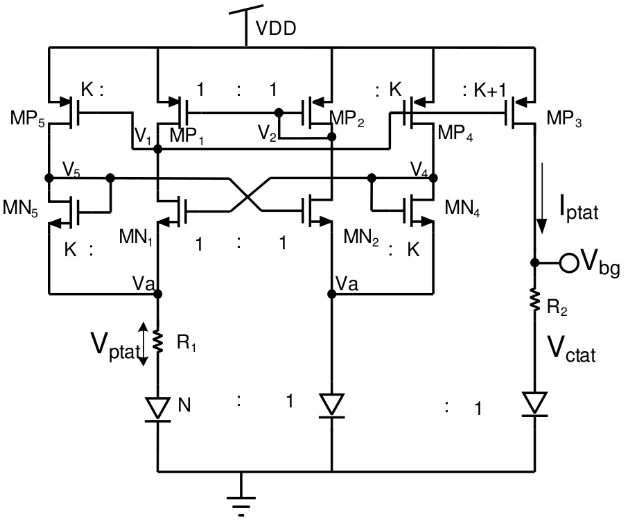


Fig. 4.20: Improved PTAT current circuit without OPA

where R_x is the dynamic resistance of both the diode-connected PMOS. When the pair transistors match well, there is $g_{m1} = g_{m2}, g_{m3} = g_{m4}, g_{m5} = g_{m6}, g_{m7} = g_{m8}$. The β_N is larger than β_P, so the system is negative feedback and therefore stable.

Setting the gain coefficient K is a trade off: increasing K can have a larger gain, but it will need more area. In this design, set $K = 15$ for compromise. In addition, most transistors are PMOS, only M_3, M_4, M_7, M_8 are NMOS. So this bandgap should have a good anti-radiation ability.

The start-up circuit is in the dotted lines, which is different with the last section. At the beginning, the gate of M_{11} is at high potential, and no current flows through R_3. So the drawn of M_{12} is low. Because of the inverter, which consists of M_{13}, M_{15}, the gate of M_{14} is high. It can switch on M_{14}, and then the gate of M_{11} changes from the VDD to the working point, at last the bandgap starts work.

After that, the R_3 voltage is high, which makes the gate of M_{14} low and switches off M_{14}. So when the bandgap works normally, the start-up circuits including $M_{12} - M_{15}$ will not work, and consume no power.

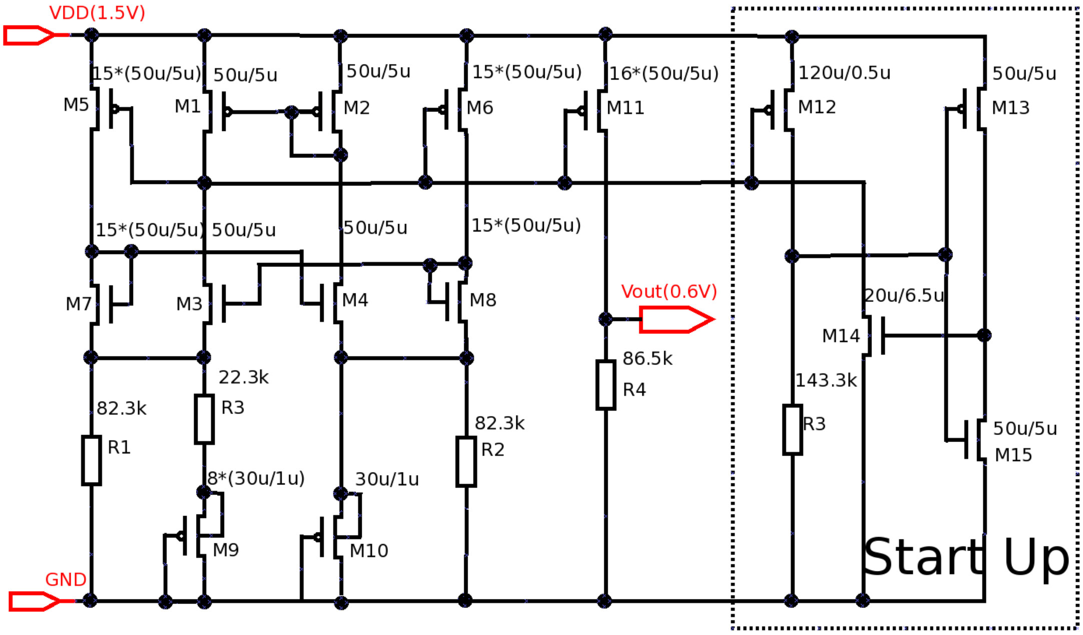


Fig. 4.21: A new architecture of bandgap without OPA

4.2.3 Radiation-Aware Design

This thesis designs two kinds of bandgaps. One uses the OPA, and the other does not use the OPA. In future, when having the chance to do the radiation test, the two bandgaps can compare each other the anti-radiation ability.

Similar with the LDOs, The main radiation effects on bandgaps are the TID and SET, and the bandgaps use the same TSMC 65 *nm* CMOS technology. So they have the better anti-TID ability. When doing the layout, all the transistor will use the guardrings, and as much as possible the substrate connectors and nwell connectors are placed in order to reduce the SET.

The parasitic triodes are prone to latch-up because of its thin base terminal. The triodes are easily influenced by radiation. So this thesis uses the diode connecting PMOS to replace the parasitic triode.

It can also simulate the radiation effect by the radiation model (200 $Mrad$(Si) and 500 $Mrad$(Si)). Here simulates the bandgap without OPA as an example. The result is shown in Figure 4.22.

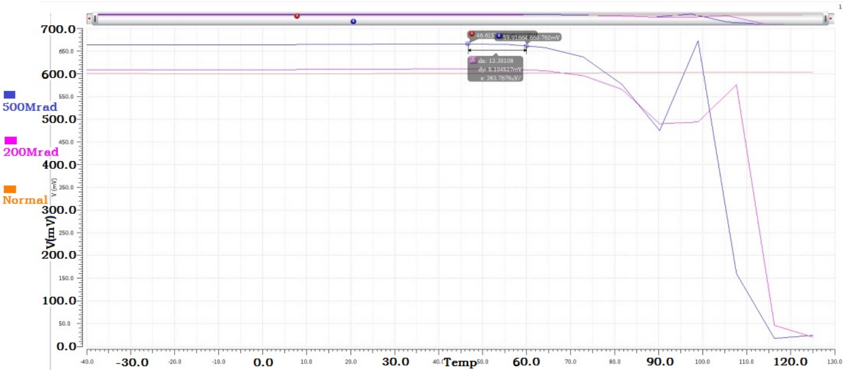


Fig. 4.22: Bandgap radiation simulation comparison

The figure shows that the work condition temperature upper limit reduces to 60 degree from 130 degree. The temperature coefficient becomes worse under 200 $Mrad$(Si) and 500 $Mrad$(Si) TID radiation, which is 38 ppm and 85 ppm respectively.

The radiation simulation shows that the TID radiation can worse the bandgap a lot under more than 200 $Mrad$(Si) TID, especially at the high temperature. But the BESIII experiment TID is much less than 200 $Mrad$(Si), and work temperature is less than 60 degree. So the bandgap can work normally in BESIII experiment. The transistors influenced by TID are NMOS, so the NMOS ELT layout can improve the result.

4.2.4 Simulation and Layout

Based on the above design, the both bandgaps are simulated. The most important bandgap parameter is the temperature coefficient in a large temperature range. In this design, the temperature range is chosen from -30°C to 130°C.

The main reason for the large temperature range is to check the bandgap whether can work normally in this range, because in some case the actual operating temperature may have such a large range. Large temperature range has an additional effect, which can check the circuit performance in the large process variation, because the temperature can change the process variation. (This check could be also achieved by other means using the available device models and the appropriate simulation tools). Not only the bandgap, other analogue circuits can use the large temperature range to check the performance variation as the process or temperature variation.

The temperature coefficient lower, the better. In this design, the output voltage of bandgap changes as temperature is shown in Figure 4.23. The left curve is the TC of bandgap without OPA, and the right one is the TC of bandgap with OPA.

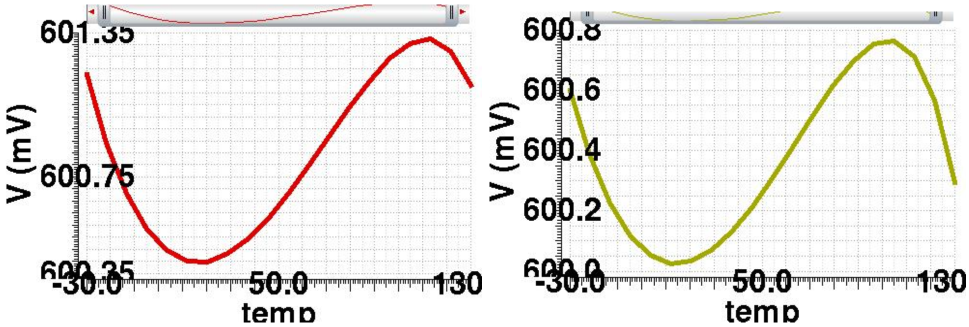


Fig. 4.23: Temperature coefficient of the bandgaps, left bandgap without OPA, right bandgap with OPA

The temperature coefficient (TC) can be expressed in the following equation:

$$\begin{aligned} TC = \frac{V_{bg,max} - V_{bg,min}}{(T_{max} - T_{min}) \cdot V_{bg,room}} \cdot 10^6 &= \frac{V_{bg,max} - V_{bg,min}}{(130 - (-30)) \cdot 600\ mV} \cdot 10^6 \\ &= \frac{V_{bg,max} - V_{bg,min}}{96\ V} \cdot 10^6 \end{aligned} \tag{4.17}$$

where $V_{bg,max}$ and $V_{bg,min}$ are the maximum output voltage and the minimum output voltage respectively, $V_{bg,room}$ is the output voltage at room temperature. T_{max} and T_{min} are the maximum and minimum temperature respectively. The TC unit is ppm, which means one millionth.

In the left curve, the output voltage changes 0.919 mV from -30 $°C$ to 130 $°C$, which means the bandgap without OPA has the temperature coefficient 9.57 ppm. In the right curve, the output voltage changes 0.728 mV in the same temperature range, so the bandgap with OPA has the temperature coefficient 7.58 ppm. The bandgap with OPA is a little better than the bandgap without OPA.

Another important feature of bandgap is PSRR, which can test the stable ability when the power supply changes. This PSRR is Power Supply Rejection Ratio, which is different definitions with the amplifier PSRR, which means Power Supply Ripple Rejection.

The PSRRs of the both bandgaps are shown in Figure 4.24.

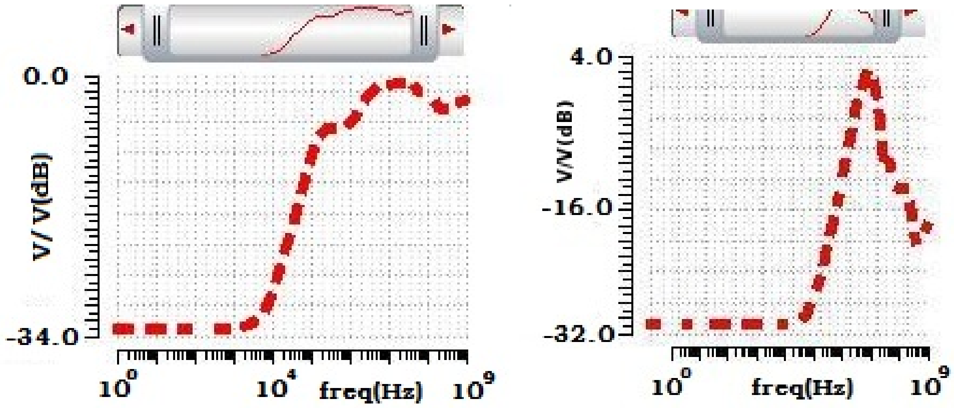


Fig. 4.24: PSRR of the bandagps, left bandgap without OPA, right bandgap with OPA

The left curve is the PSRR of bandgap without OPA, and the right one is the PSRR of bandgap with OPA. They are -34 *dB* for the bandgap without OPA and -32 *dB* for the bandgap with OPA. A good PSRR absolute value should be larger than 40 *dB*. This simulation parameter is not good with respect to 40 *dB*, because this design does not use the cascode architecture. This is one point which should be improved in future.

In order to check the mismatch influence, the Monte Carlo (MC) simulation is necessary. The MC use the random method to select the transistor parameters and then combines these random parameters to do the simulation. Before the chip goes to the factory, the MC simulation can check whether there is a high defect rate. The two bandgap PSRRs are shown in Figure 4.25.

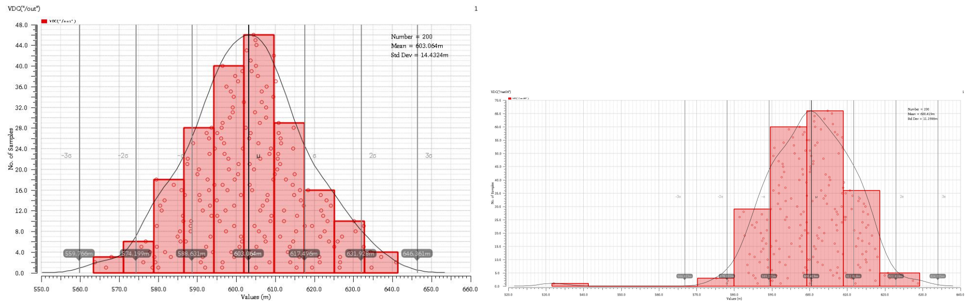


Fig. 4.25: MC simulation results of the bandagps, left: bandgap without OPA, right: bandgap with OPA

The left one comes from the bandgap without OPA, which has $\sigma = 14.4\ mV$, and 3σ spreads for 7.2%. The right PSRR is for the bandgap with OPA, which has $\sigma = 11.2\ mV$, and 3σ spreads for 5.6%. The bandgap with OPA has a better PSRR.

Although this designed bandgap power supply is 1.5 V, the bandgaps should have a good performance at the low power supply, which is useful in advanced technology. So here simulates the output voltage versus the different input voltage, which is shown in Figure 4.26.

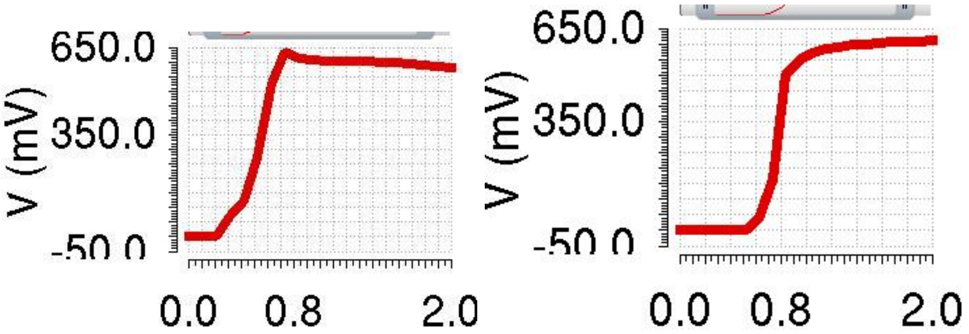


Fig. 4.26: Output voltage versus the input voltage, left bandgap without OPA, right bandgap with OPA

This figure shows that the bandgap with OPA needs a higher power supply, which is about 1.2 V; and the bandgap without OPA needs less power supply, which is about 1 V. The reason is that the OPA needs a higher power supply to work. When the power supply is 1 V or below, the bandgap without OPA will be a better choice. The line regulations can be also gotten in this figure. Through simulation, the line regulation of the bandgap without OPA is 14.4 mV/V and the bandgap with OPA line regulation is 28.6 mV/V.

When using the different supply, the temperature coefficients are also different. Figure 4.27 shows the TC at the different power supply.

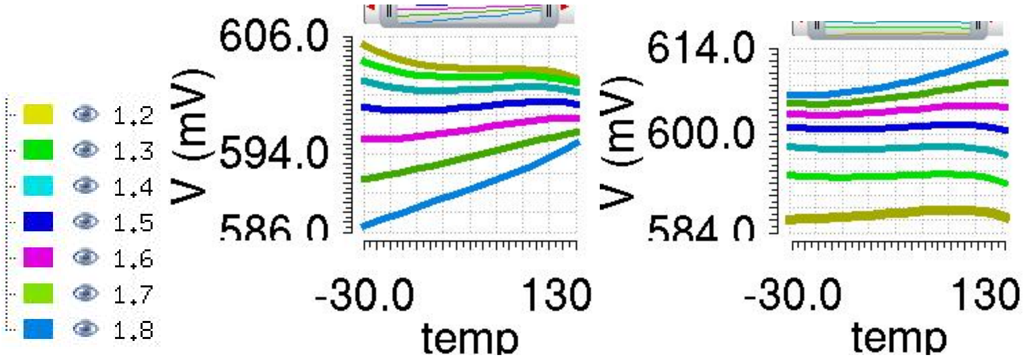


Fig. 4.27: TC at different power supplies, left bandgap without OPA, right bandgap with OPA

The different colors represents different power supplies for more clearly. The simulation results are shown in Table 4.5.

Table 4.5: Temperature coefficient at different power supply

Supply	TC of bandgap without OPA @ means voltage	TC of bandgap with OPA @ means voltage
1.2 *V*	60 *ppm*@597 *mV*	15.6 *ppm*@586 *mV*
1.3 *V*	32 *ppm*@598 *mV*	13.5 *ppm*@593 *mV*
1.4 *V*	15.8 *ppm*@599 *mV*	13.2 *ppm*@597 *mV*
1.5 *V*	7.8 *ppm*@601 *mV*	8.1 *ppm*@601 *mV*
1.6 *V*	13.5 *ppm*@602 *mV*	15.6 *ppm*@603 *mV*
1.7 *V*	23 *ppm*@603 *mV*	36 *ppm*@605 *mV*
1.8 *V*	35 *ppm*@604 *mV*	70 *ppm*@608 *mV*

The start-up circuit makes the bandgap out of the degenerate work point and starts up the bandgap. The simulation results of these circuits are shown in Figure 4.28.

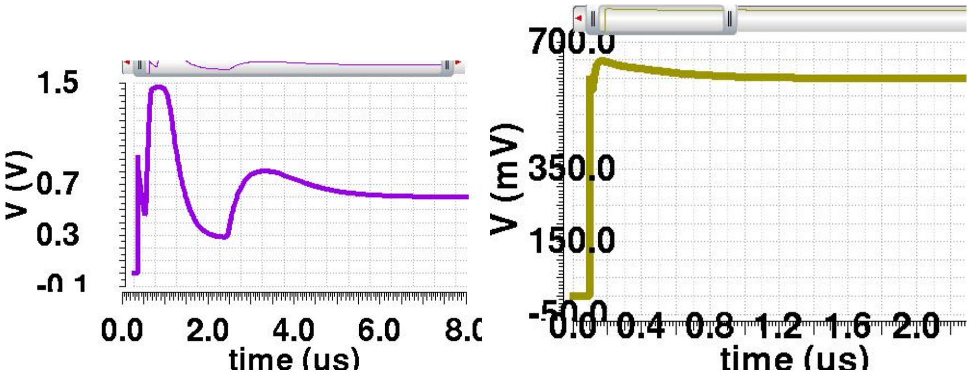


Fig. 4.28: Start-up effect of the bandagps, left bandgap without OPA, right bandgap with OPA

The right curve is better, and the start-up time of bandgap with OPA is just 800 *ns*. Bandgap without OPA needs more than 6 μs and has some shock at the beginning. The reason is the start-up circuit of the bandgap without OPA having a switch on and off actions. At the same time the bandgap without OPA needs a large transistor to reduce the mismatch, so the parasite capacitance will be large. Therefore the start-up circuit needs much time to charge and discharge the parasite capacitance. This start-up circuit has an advantage that it will not need to consume the power when the bandgap work normally.

On the other hand, the start-up circuit for the bandgap with OPA just has one switch on action, and the transistor used in the bandgap with OPA is small, so the parasitic capacitance is small. Based on these reasons, the start-up circuit needs less than 1 μs. Its draw back is that it will consume some additional current when the bandgap works normally. This two start-up circuits have different advantage and disadvantage.

The last important parameter is the noise. When added the capacitor to the output, the noise will be small. Here measures the noise without the output capacitors. The output point noise simulation result from 1 Hz to 1 GHz is shown in Figure 4.29. In the high frenquency range (larger than 1 MHz), the noise is almost zero.

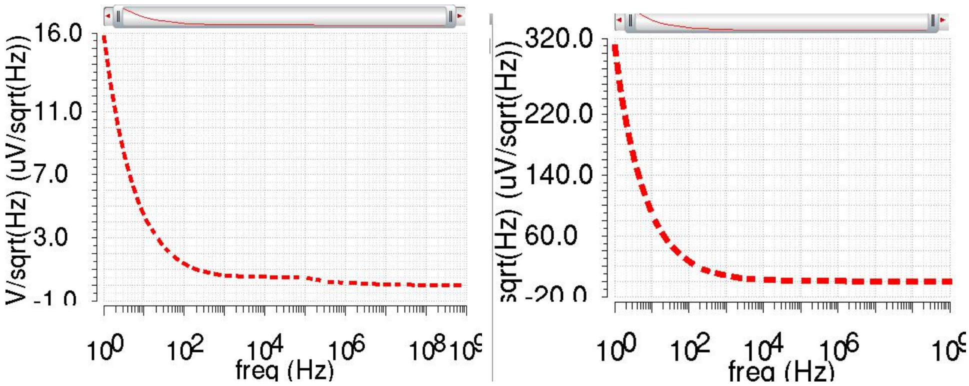


Fig. 4.29: Noise simulation result of the bandagps, left: bandgap without OPA, right: bandgap with OPA

Integrating noise from 1 Hz to 1 GHz, the noise of bandgap without OPA is 370 μV; and the noise of bandgap with OPA is 1.58 mV. The main difference comes from the OPA part. The OPA can output much noise to the output. So the bandgap with OPA has more noise appearance. To compare the both noise more clearly, detailed noise details parameters are shown in Table 4.6.

Table 4.6: Both bandgaps noise table

Frequency Hz	Noise of bandgap without OPA	Noise of bandgap with OPA
1	16 $\mu V/sqrt(Hz)$	308 $\mu V/sqrt(Hz)$
100	1.3 $\mu V/sqrt(Hz)$	25 $\mu V/sqrt(Hz)$
10 k	0.504 $\mu V/sqrt(Hz)$	2.2 $\mu V/sqrt(Hz)$
1 M	138 $nV/sqrt(Hz)$	360 $nV/sqrt(Hz)$

Both bandgaps can be used in different situations. The bandgap without OPA is better suited for applications that require low power voltage and low power consumption. The bandgap with OPA is more suitable for applications which require limited area and better spread. Table 4.7 shows the different parameters comparing between both of them, when the input is 1.5 V and output is 0.6 V.

Table 4.7: Parameters comparison between the two bandgaps

Parameters	Bandgap without OPA	Bandgap with OPA
Power	75 μW	450 μW
Minimum supply	1 V	1.2 V
Area	0.077 mm^2	0.0027 mm^2
Noise(1 to 1 GHz)	370 μV	1.58 mV
Line regulation	14.4 mV/V	28.6 mV/V
Start-up time	6 μs	800 ns
Temperature coefficient	9.57 ppm	7.58 ppm
Corner variation	23 mV	100 mV
Monte Carlo σ	14.4 mV	11.2 mV

Layout of Bandgap

For better matching, all resistors used in bandgaps are polysilicon resistors. Some of them use the dummies on both sides. All transistors connected to the pins are surrounded by guard rings.

The circuits are implemented by the TSMC 65 *nm* technology. The bandgap with OPA is shown in Figure 4.30. In order to protect this bandgap, P guard rings are used. In the layout, the diode-connected PMOS matching is the key point. The diode-connected PMOS transistors use the dummy transistors, and the two PMOS transistors have the ratio of 1:8. For a better match, set the finger number of 2:16, and the PMOS with 2 fingers will be evenly distributed in the other 16 fingers.

Another important point is the matching of $M1, M2, M5$ in Figure 4.17, the part of the arrow pointing in the figure. They affect the output current which is not sensitive to temperature. For better matching, three PMOS transistors use the same transistor, just the different figures represent the different PMOS. The two blue parts are the capacitors. One is used for the Miller capacitor, and the other is used for

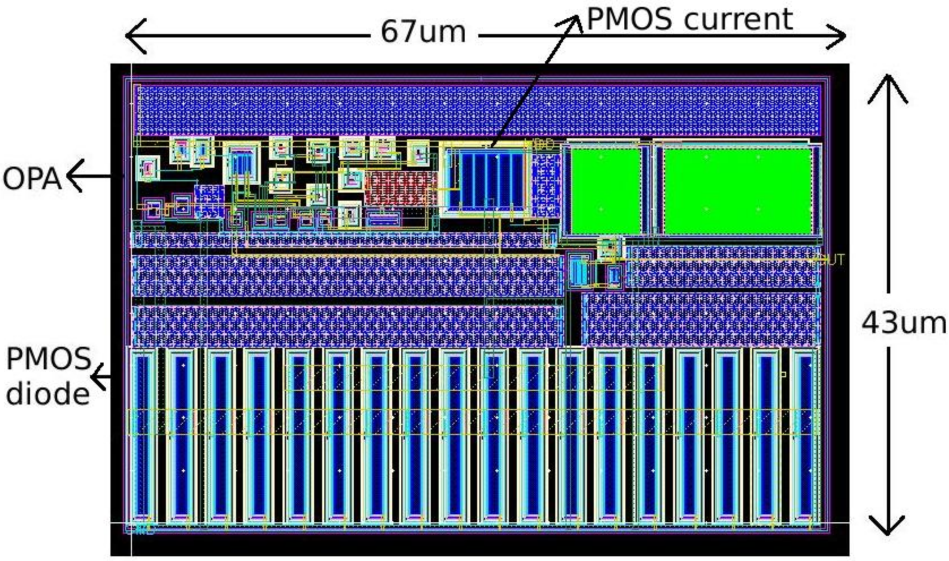


Fig. 4.30: Layout of the bandgap with OPA

stabilizing the OPA output. The small area just has the 67 μm length and 43 μm width.

Guard rings can reduce the parasitic resistance, which can cause latch-up problems, and guard rings can also reduce leakage current caused by the TID. Therefore, in order to obtain better anti-radiation capability, all PMOS transistors are surrounded by P guard rings. By the same principle, all transistors used in OPA also have guard rings.

The layout of bandgap without OPA is shown in Figure 4.31.

The two band gaps have the same part about the diode-connected PMOS. Although this bandgap without OPA is 342 μm length and 213 μm width, it occupies 0.077 mm^2. Some part of the layout is empty. The effective occupancy area is approximately 0.06 mm^2.

In the previous analysis, it is assumed that the resistance is not sensitive to temperature. But in reality, different resistors have different TCs, although they are very small. When designing a bandgap with very small TC, the resistance TC must be considered.

Fortunately, some resistors have opposite temperature coefficients. To make the resistor temperature insensitive, the resistor can be implemented with complementary

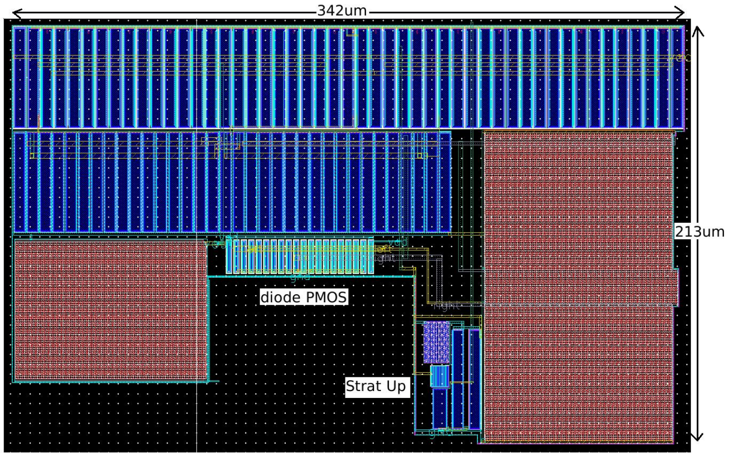


Fig. 4.31: Layout of the bandgap without OPA

TC resistors such as N+ Ploy and P+ Poly resistors. Careful design of the ratio of the two series resistors gives a temperature-insensitive resistor.

4.2.5 Conclusions

Low power consumption and low output voltage bandgap will find more applications in advanced technologies, especially the 180 *nm* or below. This chapter designs two kinds of bandgaps using some radiation aware technologies. Although the both bandgaps perform basic functions and get a very small temperature coefficient, they needs to be improved on some points.

One point is PSRR, which needs to be increased. PSRR is more important when bandgap is used in hybrid circuits. Digital circuits always affect the power supply. A lower PSRR introduces larger power supply noise into the analogue portion. One method is to use a cascode circuits at the output.

Another point is the bandgap without OPA occupying large area. There should be some technology to reduce the area. Small area bandgap are highly competitive in advanced technologies, especially when bandgaps are distributed. In this case, the bandgap not only has a few of power consumption but also has a small area. A small distributed bandgap along with a distributed LDO will provide a better power supply for multichannel front-end circuits.

The last one is the limited output current, This parameter is weak and should be strengthened. Although the bandgap is primarily used to provide a reference voltage to some gates of the transistors. As the technology node goes down below 130 *nm*, the quantum tunneling effect will play a role, which causes the gate leakage current to increase a lot. So advanced technology requires the bandgap to have enough current drive capability.

Finally, comparing the both bandgaps with other paper's works, the results are shown in Table 4.8. As can be seen from the table, the band gaps of both designs have a low temperature coefficient. The bandgap with OPA takes up almost the smallest area. This table shows that the work of this thesis is advanced in some parameters, but some parameters are mid-range performance and need to be improved in the future.

Table 4.8: Performance comparison with others bandgaps

Parameters	Technology	Diode Type	Minimum supply	Output Voltage	Current	Temp Range	Temp Co-efficient	Spread (3σ)	Active area
Panchanan [32]	0.16 μm	DTMOST	0.9 V	670 mV	2.4 μA	-40 to 125	50 ppm	1.25%	0.05 mm^2
Panchanan [32]	40 nm	DTMOST	0.8 V	571 mV	2.58 μA	-40 to 125	34 ppm	3.7%	0.05 mm^2
Ge[33]	0.16 μm	BJT	1.7 V	1.0875 V	55 μA	-40 to 125	12 ppm	0.75%	0.12 mm^2
Lam[34]	0.35 μm	BJT	0.9 V	635 mV	16.6 μA	5 to 95	24 ppm	n/a	0.059 mm^2
Annema[35]	0.35 μm	DTMOST	0.85 V	650 mV	1.2 μA	-20 to 100	60 ppm	6%	0.063 mm^2
Annema[36]	0.16 μm	MOSFET	1.1 V	944 mV	1.5 μA	-11 to 135	30 ppm	2.25%	0.0025 mm^2
Annema[37]	32 nm	FINFET	0.9 V	600 mV	14 μA	0 to 150	962 ppm	n/a	0.016 mm^2
Vita[38]	0.35 μm	MOSFET	0.9 V	670 mV	40 nA	0 to 80	10 ppm	9.3%	0.045 mm^2
Giustolisi[39]	1.2 μm	MOSFET	1.2 V	259.3 mV	3.58 μA	-25 to 120	154 ppm	3.66%	0.23 mm^2
Kinget[40]	90 nm	MOSFET	0.55 V	247 mV	870 μA	-25 to 120	150 ppm	7.45%	0.07 mm^2
This work with OPA	65 nm	MOSFET	1.2 V	600 mV	50 μA	-30 to 130	7.58 ppm	5.6%	0.0027 mm^2
This work without OPA	65 nm	MOSFET	1 V	600 mV	300 μA	-30 to 130	9.57 ppm	7.2%	0.077 mm^2

Chapter 5

The Transistor Design of the BUCK IP Block Using 65 *nm* CMOS Technology

5.1 Introduction

There are two types of switches DC-DC in the normal power supply. One is an isolated DC-DC, which often uses a transformer as a power converter. This DC-DC is typically used in high power applications. Due to the isolation, the second side has a small effect on the first side and is therefore safer. In addition to large size, transformers are difficult to manufacture using CMOS technology. This isolated DC-DC is difficult to integrate in front-end design.

The other is non-isolated DC-DC. They fall into two categories: switched capacitors and switches LC. The switched capacitor DC-DC implements a voltage regulator by charging and discharging the capacitor.

Both configurations can produce boost or buck. The charge pump can even output a reverse voltage and it does not use an inductor. However, the effect is less than the switch LC, and the output power is limited. It is especially suitable for small current and high voltage conditions such as PMT power supplies. PMT even requires several thousand volts of current, but the current is less than 1 *mA*.

This design uses an inductor as a power transfer to output a large current with high efficiency. Inductors require a large area in CMOS technology. Therefore, the inductor is usually placed outside the chip or the bonding coil is used as an inductor. In order to improve the efficiency of the inductor, several switches DC-DC can share one inductor.

According to the relationship between the output voltage and the input voltage, the switch LC can be divided into three categories: [41], as shown in Figure 5.1. One is the buck type (a), which means that the output voltage will be lower than the input voltage; the other is the boost type (b), which means that the output voltage is higher than the input voltage; the third is the buck-boost type (c), which means that the output voltage can be higher or lower than the input voltage.

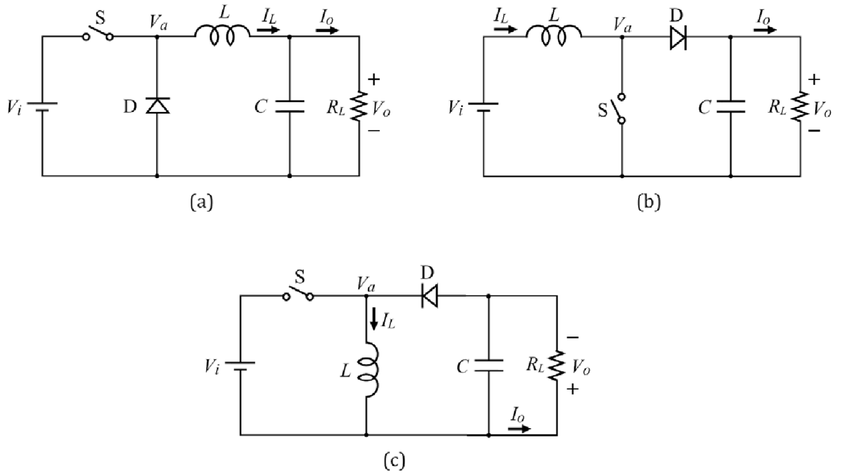


Fig. 5.1: Typical DC-DC converters

Typically in front-end designs, the ASIC's voltage is less than the board's power supply, so the switching power supply DC-DC is typically used to lower the board power. The buck DC-DC can be used as a first stage converter, which can be followed by an LDO.

5.2 Architecture of the BUCK

A simple buck conversion is shown in Figure 5.1 (a), which consists of a switch, a diode, an inductor, a capacitor and a resistive load. The switch has a stable frequency. The relationship between the switch and V_a is shown in Figure 5.2.

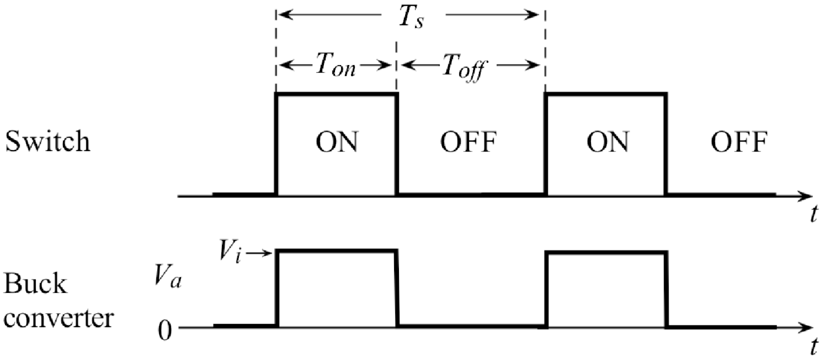


Fig. 5.2: Relationship between the switch and V_a

where T_s is the frequency period, T_{on} and T_{off} are the switch ON and OFF time, respectively. T_{on}/T_s is an important parameter, which is defined as the duty cycle D:

$$D = \frac{T_{on}}{T_{on} + T_{off}} = \frac{T_{on}}{T_s} \tag{5.1}$$

When the switch is turned on, the inductor current will rise, but since the inductance characteristic slows down the current change, the process is slow. By the same principle, when the switch is turned off, current will still flow through the diode, but the rate of current drop is small. Therefore, the buck circuit can output a continuous current to the load, and the function of the capacitor is to reduce the output voltage ripple.

When the buck converter is in a steady state, the rising current of each circle is equal to the falling current. Therefore, the current integration through the inductor is zero. This is the voltage quadratic balance. It can be derived as follows:

$$V = L \cdot \frac{\mathrm{d}i}{\mathrm{d}t} \tag{5.2}$$

where V is the voltage drop of the inductor; i is the inductor transient current; t is the time; and L is the inductor inductance. This equation can be transferred as follows:

$$V = L \cdot \frac{\Delta i}{\Delta t} \tag{5.3}$$

$$\Delta i = \int_0^{\Delta t} \frac{V}{L} \mathrm{d}t = \frac{V \cdot \Delta t}{L} \tag{5.4}$$

When the voltage second balance is applied to the buck converter, the switch on state Δi equal to the off state Δi. During the on state, the inductor voltage is $V_i - V_o$, and in the off state, the inductor voltage is $-V_o$. So there are the following equations:

$$\frac{V_i - V_o}{L} T_{on} = \frac{V_o}{L} T_{off} \tag{5.5}$$

$$\frac{V_o}{V_i} = \frac{T_{on}}{T_{on} + T_{off}} = D \tag{5.6}$$

$$V_o = V_i \cdot D \tag{5.7}$$

Considering the efficiency of the converter η:

$$V_o = V_i \cdot D \cdot \eta \tag{5.8}$$

From the equation above, the output depends on the duty cycle D. Adjusting the duty cycle can change the output voltage. The negative feedback section can sample the output voltage, control the duty cycle, and ultimately bring the output voltage to the desired value.

According to the waveform of the inductor current, there are two types: continuous conduction mode (CCM) and discontinuous conduction mode (DCM). Based on the feedback control method, the buck converter can be divided into voltage mode control (VMC) and current mode control (CMC). All of the different buck types described above can be modulated by pulse width modulation (PWM) or pulse frequency modulation (PFM) techniques.

5.2.1 CCM and DCM Mode

The CCM mode means that the inductor current will not be zero for one switching cycle, so the inductor will never reset. Even if the power MOSFET is turned off, the inductor's inductor flux will not return to zero. The DCM mode is reversed. During the switching cycle, the inductor current will be zero and the inductor will reset. When the output current is large, the buck converter usually works in CCM mode, as shown in Figure 5.3.

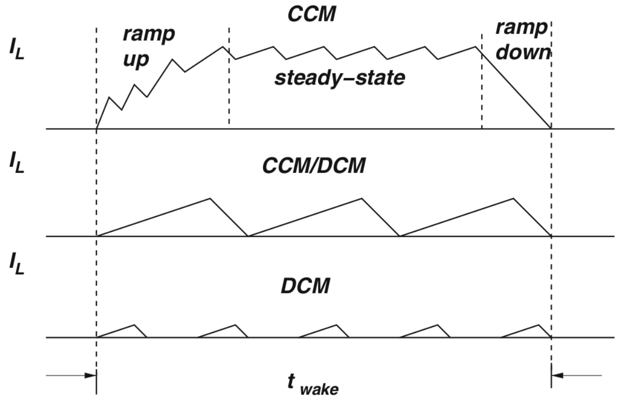


Fig. 5.3: CCM and DCM modes

As the output current decreases, the operating mode will change from CCM mode to DCM mode. In DCM mode, the power efficiency is higher and the current ripple will be larger. In order to keep the output voltage stable, the duty cycle must vary depending on the output current. In this design, the CCM mode is superior to the DCM mode. To maintain the CCM mode, the output inductor can be amplified to reduce the critical threshold output current.

5.2.2 PWM and PFM Mode

All different buck types can be modulated by PWM or PFM techniques. The PWM regulates the output voltage and changes the duty cycle with the same switching frequency. It is shown in Figure 5.4.

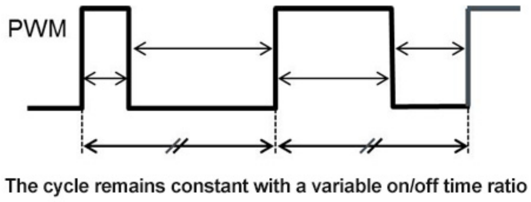

PWM
The cycle remains constant with a variable on/off time ratio

Fig. 5.4: PWM regulate mode

Due to the fixed frequency, the output noise caused by the switch is fixed, so noise can be more easily reduced by a simple RC filter. At the same time, the power lost through the switching power transistors is about the same whether the load is heavier or lighter.

The power loss of the power transistor depends on the switching frequency. Therefore, there is a trade-off between power efficiency and output inductance. At the same ripple level, higher frequencies require lower inductance. Therefore, the output inductor is smaller at the expense of lower power efficiency. And when the load current is small, the power efficiency is also low.

In PFM mode, the on (or off) state can remain exactly the same and change the off (or on) state time for different currents. So the frequency will change accordingly. The cycle frequency will vary depending on the load current. It is shown in Figure 5.5.

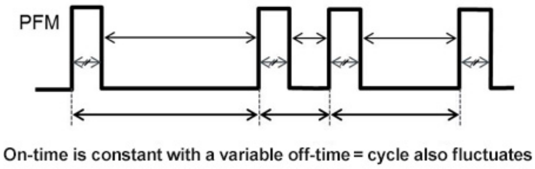

PFM
On-time is constant with a variable off-time = cycle also fluctuates

Fig. 5.5: PFM regulate mode

The figure above shows that the on state keeps the same time. It adjusts the off state time to accommodate the load current. When the load current is large, the off state time is shortened. When the load current is small, the off state takes a long time. Therefore the frequency is in a variable state and is inversely proportional to the load current.

Due to the variable frequency, the filter is difficult to design. But this PFM is more suitable for light load currents. When the load current is small, the switching frequency is small, so the power efficiency is high.

5.2.3 VMC and CMC Feedback Mode

According to the feedback control method, the buck converter can be divided into VMC and CMC [42]. The purpose of the two modes is to control the PWM output. The difference is that the CMC uses the output current for negative feedback; and the VMC uses the output voltage for negative feedback.

The typical structure of CMC is shown in 5.6. The current mode has two feedbacks. One is the current feedback used to control the inductor current and the other is the voltage feedback used to control the output voltage.

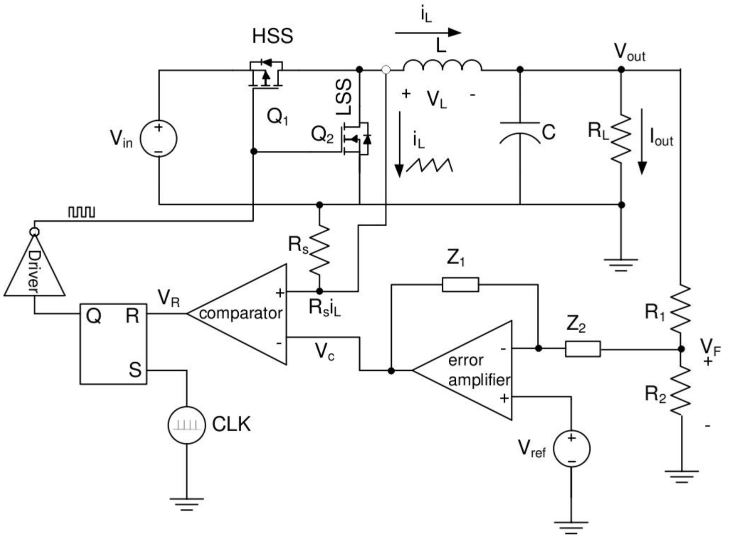


Fig. 5.6: Current control mode structure

The sample current from the inductor is converted to voltage R_si_L by the sampling resistor R_s. The voltage is then compared to the output of the error amplifier V_c by a comparator. In a cycle, V_c is almost the same.

R_si_L voltage is connected to the non-inverting input of the comparator, and V_c is connected to the inverting input. The output of the comparator is connected to the R terminal of the R-S latch. The truth table for R-S latch is shown in Table 5.1.

Table 5.1: Truth table of the S-R latch

S	R	Q	$\overline{Q}$
0	0	latch	latch
0	1	0	1
1	0	1	0
1	1	0	0

When the inductor current increases to a certain value, R_si_L will be larger than V_c. Therefore, the comparator will output a zero voltage to the R terminal of the S-R latch and reset the S-R latch. Output Q will produce a low voltage level. Then S and R are 0 and 1, respectively, and the output Q is 0, turning off the power transistor. It reduces the current flowing through the inductor.

When the current drops below V_c, the comparator gives the R terminal 0. Together with 0 at the S terminal, the latch outputs a latched state. It outputs the last state 0 to the power transistor. Therefore, the inductor current will continue to drop until the end of this cycle. At this time, S is reset to 1. At the R terminal together with 0, the output Q will become 1, thereby turning on the power transistor. Even if the S terminal returns 0 after a very short pulse, the 0 of S and R locks the last state 1. So it will continue to expand the inductor current until $R_S i_L$ exceeds V_c again.

The voltage control method has only one feedback, as shown in Figure 5.7.

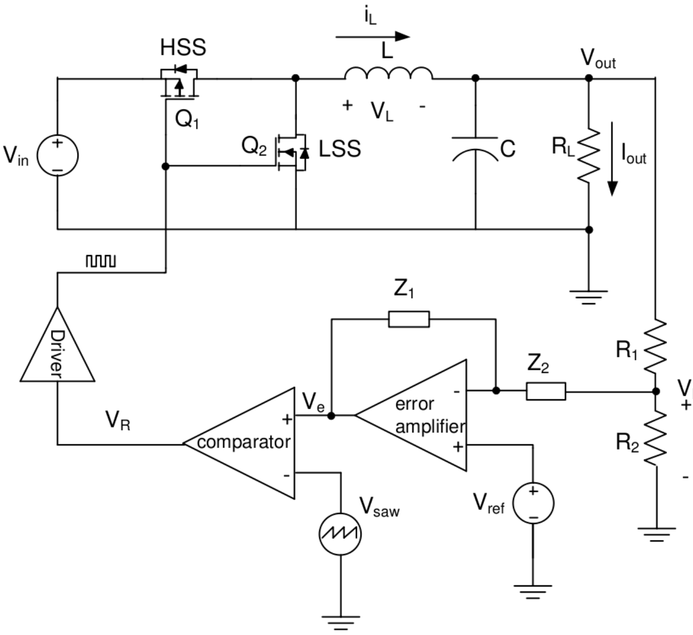


Fig. 5.7: Voltage control mode structure

In voltage mode, the feedback sampled signal is the output voltage, sampled by R_1 and R_2. The divided voltage is applied to the inverting terminal of the error amplifier. The difference between this voltage and the reference voltage at the non-inverting terminal will be amplified by the error amplifier.

The amplified signal will be placed on the non-inverting side of the comparator and compared to the ramp saw signal, which is at the inverting end. The output of the comparator will output a modifiable pulse period wave with the same frequency as the ramp saw signal. The pulse period wave is a PWM signal that can control the ON or OFF of the power switching transistor.

When the output voltage drops below the required voltage, the sampled voltage will decrease. When the output of the error amplifier V_e decreases, the comparator will output a PWM signal with a shorter ON state pulse. The shorter ON state pulse will reduce the power-on time, which will reduce the output current and thus the output voltage.

The voltage feedback mode uses only one voltage feedback. Therefore, comparing current feedback requires a more complex compensation circuit to keep the system stable. But the current feedback has some drawbacks. Not only is it more complicated, but current feedback also introduces some current noise into the feedback path, so sometimes the current feedback noise is greater than the voltage feedback noise.

Taking into account the above analysis, the buck switch DC-DC usually adopts CCM mode, PWM modulation and VMC feedback mode. These combination modes are more suitable for high current loads with better noise performance.

5.3 Transistor-Level Design of the BUCK

According to the previous section, the design of the buck switch DC-DC is shown in Figure 5.8. The buck DC-DC uses voltage feedback control and PWM modulation. The main mode of operation is the continuous conduction mode.

The buck switch DC-DC consists of three modules. One is the converter power block, the other is the compensator block, and the other is the modulator block. The modulator block and compensator block consist of feedback control that controls the output at a regulated voltage.

The buck converter converts the input voltage (2.5 V to 3.5 V) to 1.5 V to power the LDO. The frequency is 1 MHz, the maximum current is 1 A, the minimum output current is 5 mA, and the voltage ripple is less than 5 mV.

Due to the limited area and pads, some components are removed off-chip, including power switching transistors, inductors, and capacitors. These components can be purchased from commercial companies. The rest of the buck architecture will be on the chip. On-chip and off-chip components are shown in Figure 5.9. The red dashed line is the on-chip portion and includes most of the buck architecture.

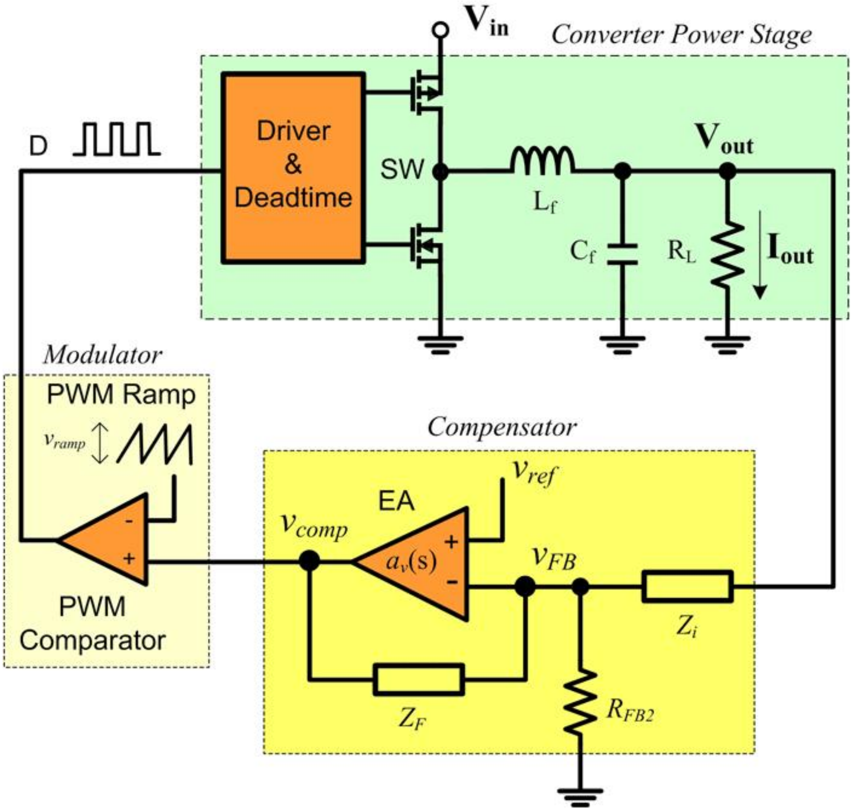


Fig. 5.8: Simple buck architecture mode

5.3.1 Converter Power Stage

The converter power stage block consists of three parts. One is the driver and dead time controller, the other is the power switching transistor (PMOS and NMOS), and the third is the inductor and capacitor for energy transfer.

The driver and dead time controller play two roles. One is to extend the drive current and voltage capability. Power switching transistors typically have larger sizes to reduce on-resistance, thereby reducing power consumption. Some switching transistor powers have a linear relationship with the transistor on-resistance.

Another role is to control the switching states of the PMOS and NMOS. The purpose is to prevent the power switching transistors from being turned on at the same time. Otherwise, the input voltage and ground may form a short circuit, which increases the power consumption of the converter.

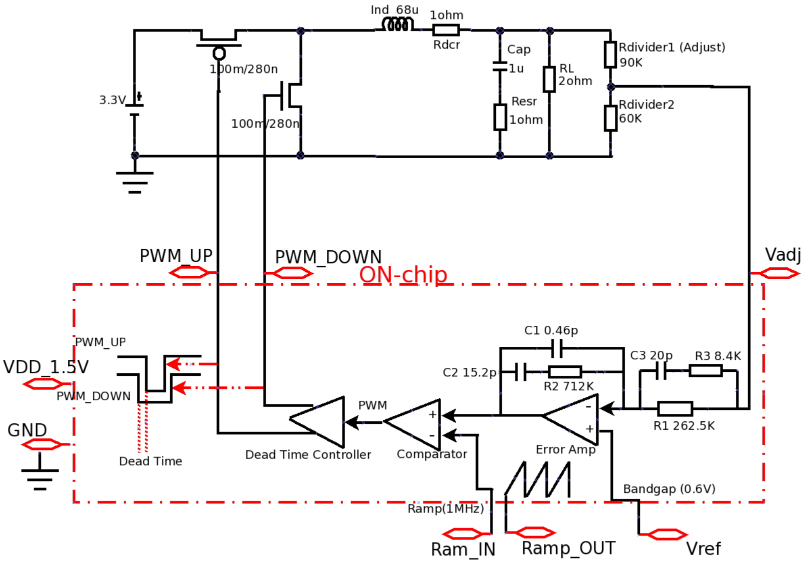


Fig. 5.9: On-chip and off-chip components distribution

Power transistors include PMOS and NMOS. PMOS replaces the switch in Figure 5.1 (a).The NMOS replaces the diode.

Asynchronous and Synchronous

The diode is a unidirectional switch, which makes the system asynchronous because the forward conduction will consume some voltage. Due to the unidirectional conduction of the diode, the system will convert from CCM to DCM when the current is low. It is shown in Figure 5.10.

The MOSFET is bidirectional. Therefore, when MOS is used instead of a diode, the voltage drop across the transistor is small. This synchronizes the system and consumes less power. It is shown in Figure 5.11.

As can be seen from the two figures, the synchronous and asynchronous systems are in CCM mode when the current is large. However, when the output current is small, the asynchronous mode is in DCM mode, and the synchronization is in CCM mode. Since the designed buck converter will be used for the 0.6 A output, the design chooses a synchronous system to increase efficiency. Although it can bring dead time and reduce efficiency at low current.

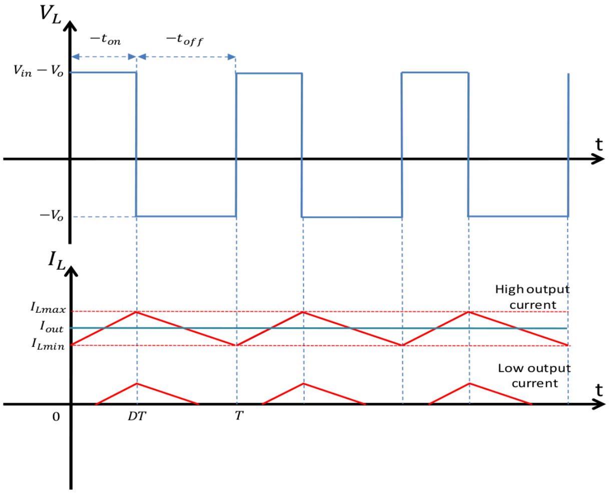


Fig. 5.10: Diode asynchronous buck inductor currents

In order to keep the system in CCM mode, the inductor should be carefully selected. The CCM and DCM boundaries should be studied carefully, as shown in Figure 5.12.

where L is the inductance; V_{imax} is the maximum input voltage; V_{imim} is the minimum input voltage; I_{OB} is the average output current; DT is the switch on state time; T is the frequency cycle. In order to maintain CCM, the minimum current should be:

$$I_{min} = \frac{1}{2}\Delta i_{Lmin} \tag{5.9}$$

Inductance Parameter

CCM or DCM mode, the current ripple, the voltage ripple and some other important parameters can be influenced by the inductor and the capacitor, so they should be chosen firstly [42]. According to the nature of the inductance, it can be gotten:

$$V_L = L\frac{dI_L}{dt} \tag{5.10}$$

For $0 < t \leqslant DT$:

$$i_L = \frac{V_I - V_O}{L}t \tag{5.11}$$

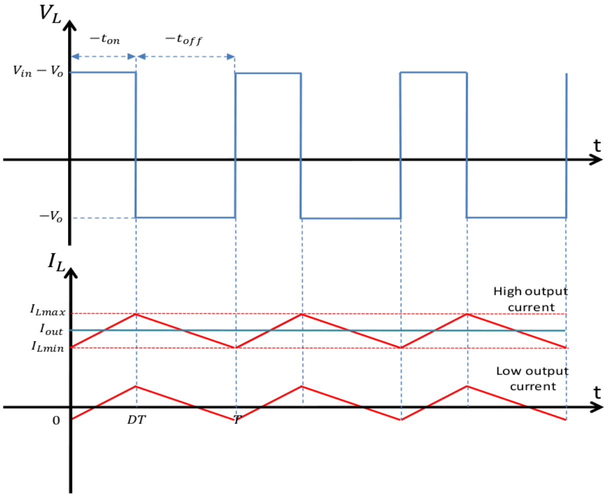


Fig. 5.11: MOSFET synchronous buck inductor currents

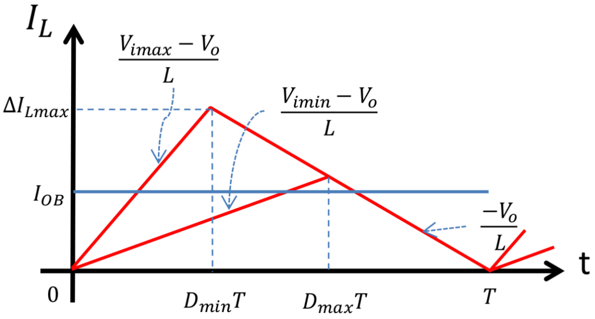


Fig. 5.12: Current at the CCM/DCM boundary

where V_i is the input; V_o is the output; L is the inductance. So in Figure 5.12, the current maximum rising slope is $(V_{imax}-V_o)/L$. The current minimum rising slope is $(V_{imin}-V_o)/L$. When the switch is turned off, the inductor current will decrease, ie:

$$i_L = \frac{0-V_O}{L}t = -\frac{V_o}{L}t \tag{5.12}$$

The down slope is always $-V_o/L$, as shown in Figure 5.12. The peak current at this boundary can be solved as follows:

$$\Delta i_L = i_L(DT) = \frac{V_i - V_o}{L}\cdot D\cdot T = \frac{V_o(1-D)}{L\cdot f_s} \tag{5.13}$$

$$\Delta i_{Lmax} = \frac{V_o(1-D_{min})}{L \cdot f_s} \tag{5.14}$$

$$\Delta i_{Lmin} = \frac{V_o(1-D_{max})}{L \cdot f_s} \tag{5.15}$$

$$L = \frac{V_o(1-D)}{\Delta i_L \cdot f_s} = \frac{V_o(1-D)}{2 \cdot I_{OB} \cdot f_s} \tag{5.16}$$

As can be seen from the equation (5.16), the inductance is inversely proportional to the frequency. The higher the frequency, the less the inductance and the smaller the inductor area. But a larger switching frequency means higher switching power consumption, which reduces efficiency. This is a trade-off. This design chooses a frequency of 1 *MHz* for tradeoffs.

On the boundary, Δi_L is 2 times the I_{OB}. The current ripple Δi_L and the minimum output current I_{OB} are also inversely proportional. The larger inductance means the fewer current ripple and the fewer output current when working in the CCM mode. In order to maintain a large application range, the output current ranges from 5 *mA* to 1 *A*. Therefore the minimum output current is 5 *mA*.

The input voltage range is from 2.5 *V* to 3.3 *V*, which is typically used as the voltage for the board's power supply. When the output is 1.5 *V*, the duty cycle ranges from 0.45 to 0.6. The maximum duty cycle gives the minimum inductance. Based on these parameters, the minimum inductance can be gotten:

$$L_{min} = \frac{V_o(1-D_{min})}{2 \cdot I_{OBmin} \cdot f_s} = \frac{1.5\ V \cdot \left(1 - \frac{1.5\ V}{3.3\ V}\right)}{2 \cdot 5\ mA \cdot 1\ MHz} = 41\ \mu H \tag{5.17}$$

For nominal commercial inductance values, there is no 41 μH value. The standard commercial value above this one is 68 μH. So here chooses 68 μH. The next step is to choose the minimum output capacitance based on the request voltage ripple and inductance.

Capacitance Parameter

The minimum capacitance depends on the minimum voltage ripple requested. Referring to Figure 5.13, the output capacitor has an equivalent series parameter resistance

(ESR) r_C. The inductor current flows through the shunt load resistor R_L and the output capacitor C with ESR.

The best effect is all direct current (DC) flowing through R_L and all alternating current (AC) flowing through the capacitor. In that ideal case, there is no voltage ripple. The reality is that most of the AC current flows through the capacitor, with very little of it flowing through the load resistor. The latter forms a voltage ripple.

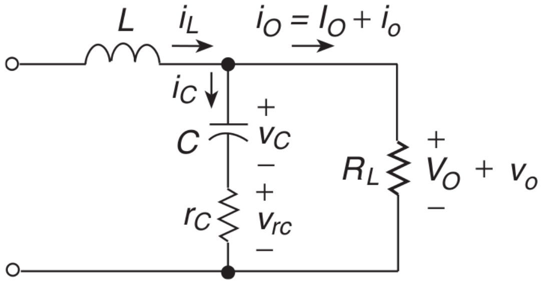


Fig. 5.13: Output capacitor with ESR

The AC current flowing through the capacitor forms two voltages: the capacitor voltage V_C and the ESR voltage V_rc. The current flowing through the capacitor, i_c, is shown in Figure 5.14.

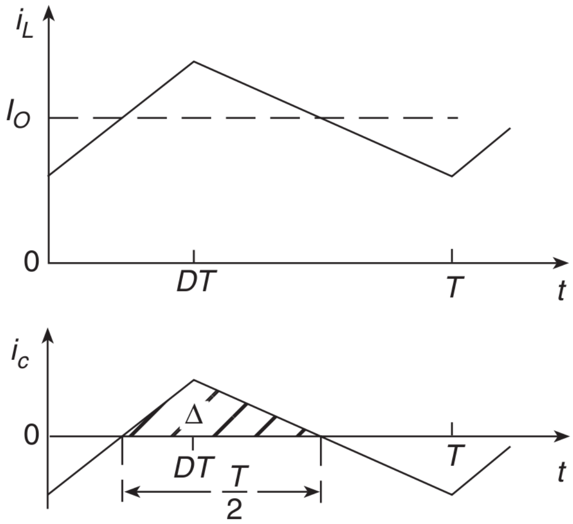


Fig. 5.14: AC current of the output capacitor

The AC inductor current can be obtained by subtracting the DC current from the inductor current. The AC current can be seen as the CCM / DCM boundary current, so the equation from CCM/DCM can be used for AC current.

$$i_c = i_L - Io \tag{5.18}$$

The ESR voltage V_rc is shown in Figure 5.15. The shape is the same as the AC current. And the peak to peak ESR voltage V_{rcpp} is the maximum current Δi_{Lmax}. Referring to the equation (5.14), multiply the ESR resistance r_C:

$$V_{rcpp} = r_C \Delta i_{Lmax} = \frac{r_c V_o(1 - D_{min})}{f_s L} \tag{5.19}$$

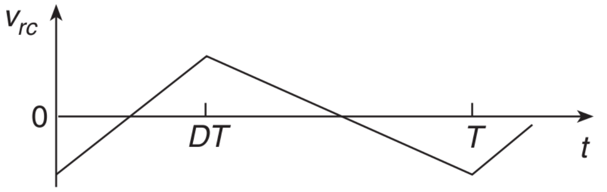


Fig. 5.15: ESR resistor voltage

The voltage across the capacitor is caused by the AC inductor current. In Figure 5.14, the shaded area indicates that the AC current is greater than zero. During this time, charge flows into the capacitor and causes the capacitor voltage to rise. In other parts of the cycle, AC current flows out of the capacitor and causes the capacitor voltage to drop. Therefore, the change in the current in the shaded area causes a change in the peak-to-peak value of the output voltage. It is shown in the graph 5.16.

The shaded area represents the accumulated charge of the inductor current. It can be calculated:

$$\Delta Q = \frac{1}{2} \cdot \frac{T}{2} \cdot \frac{\Delta i_{Lmax}}{2} = \frac{1}{2} \cdot \frac{T}{2} \cdot \frac{1}{2} \frac{V_o(1 - D_{min})}{f_s L} = \frac{V_o(1 - D_{min})}{8 L f_s^2} \tag{5.20}$$

The peak to peak capacitor voltage can be gotten:

$$V_{Cpp} = \frac{\Delta Q}{C} = \frac{V_o(1 - D_{min})}{8 L f_s^2 C} \tag{5.21}$$

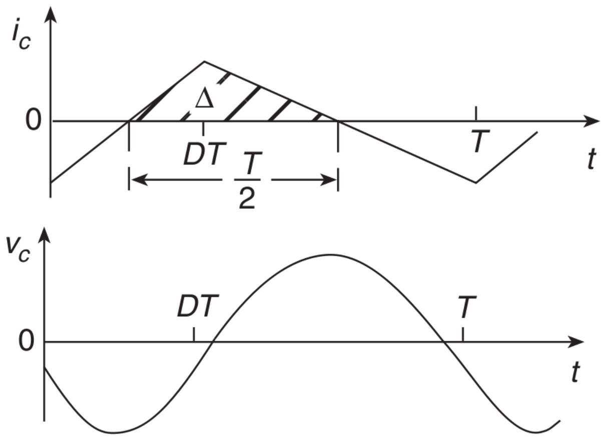


Fig. 5.16: Output capacitor voltage variation caused by the AC current

The ESR voltage phase is different from the capacitor voltage phase, so the overall voltage ripple is approximately equal to the sum of the absolute values of the two voltages.

$$V_r \approx V_{Cpp} + V_{rcpp} = \frac{V_o(1-D_{min})}{8f_s^2LC} + \frac{r_C V_o(1-D_{min})}{f_s L} \tag{5.22}$$

The nominal ESR is approximately 200 $m\Omega$. So there are the following parameters:

$$\begin{cases} f_s = 1\ MHz; \\ L = 68\ \mu H; \\ V_o = 1.5\ V; \\ D_{min} = \frac{1.5\ V}{3.3\ V} = 0.455 \\ r_C = 200\ m\Omega; \end{cases} \tag{5.23}$$

The following equation can be gotten:

$$V_r \approx \frac{1.5 \times 10^{-9}\mathrm{C}}{C} + 2.4\ mV \tag{5.24}$$

The output voltage ripple is inversely proportional to the capacitance and is proportional to the ESR. In order for V_r to be less than 5 mV, the capacitance should be greater than 0.58 μF, so the minimum output capacitance is 0.58 μF.

If the ESR is slightly greater than 200 Ω, then the output capacitor is chosen to be 1 μF. From this equation, a higher frequency can reduce the capacitance value, thereby reducing the size, just like the output inductor. And higher frequencies increase switching power consumption. So this is a trade-off.

Power Loss and the MOSFET Parameters

One of the advantages of switching DC-DC is its high efficiency. Most of the power loss is generated in the power stage block. The main power loss list is as follows:

1. The conduction loss of the output inductor P_{IND};
2. The ESR loss of the output capacitor P_{ESR};
3. The conduction loss result from the resistance of the MOSFETs P_{CONH},P_{CONL};
4. The switch loss of the MOSFETs P_{SWH}, P_{SWL};
5. The reverse recovery loss caused by the body diode P_{DIODE};
6. The parasitic capacitance loss in the MOSFETs P_{COSS};
7. The parasitic capacitance loss on the gate P_{GATE}
8. The dead time loss P_{DEAD};

Now, when the output current is set to 1 A , the following section will calculate each power loss and achieve overall efficiency.

Conduction Loss of the Inductor The power loss caused by the inductor comes from two parts. One is the inductor core loss caused by magnetism. The inductor core loss is derived from the alternating magnetic field in the core material. It depends on the frequency and total flux swing. This part of the computation equation is very complicated, and this power loss is small and therefore negligible.

The other is caused by the parasitic resistance of the inductor and the resistance of the wire. Wire resistance is usually small to negligible compared to inductor parasitic resistance.

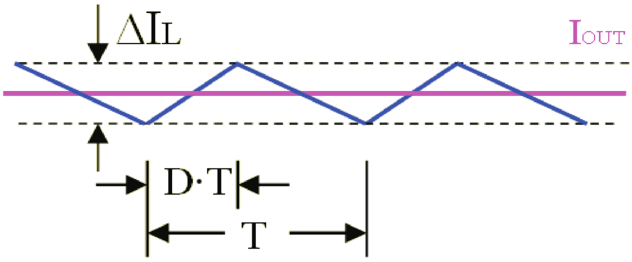


Fig. 5.17: Current variation of the inductor

The inductor current is shown in Figure 5.17. The inductor current consists of the DC current I_{OUT} and the AC current. P_{IND} is given by:

$$P_{IND} = I^2_{RMS\_L} \cdot R_{DCR} \tag{5.25}$$

where R_{DCR} is the parasitic DC resistance of the inductor, and I_{RMS_L} is the effective current value of the inductor.

$$I^2_{RMS\_L} = I^2_O + \frac{\Delta I^2}{12} \tag{5.26}$$

where ΔI is the ripple current, typically less than 30% output current. In this design the ΔI is 0.5% of the output current (1 A). So it can be neglected.

The inductor uses TDK's VLP8040 with an inductance of 68 a parasitic resistance of 0.19 Ω. Therefore, when the output current is 1 A, the inductor power loss is:

$$P_{IND} = I^2_{RMS\_L} \cdot R_{DCR} \approx 1\ A \cdot 1\ A \cdot 0.19\ \Omega = 0.19\ W \tag{5.27}$$

Conduction Loss of the Output Capacitor The power loss of the output capacitor comes from the parasitic resistance ESR. ESR includes not only parasitic series resistance but also equivalent resistance from capacitor leakage current and dielectric loss.

Since most of the AC current flows through the ESR, the ESR power loss is:

$$P_{ESR} = I^2_{RMSAC} \cdot R_{ESR} \tag{5.28}$$

where I_{RMSAC} is the effective current, which is $\Delta I_L^2/12$, so there is the following equation:

$$P_{ESR} = \frac{\Delta I_L^2}{12} \cdot R_{ESR} \tag{5.29}$$

$$\Delta I_L = \frac{V_{IN} - V_{OUT}}{f_{SW} \cdot L} \cdot \frac{V_{OUT}}{V_{IN}} \tag{5.30}$$

where V_{IN} is the input voltage; V_{OUT} is the output voltage; f_{SW} is the switching frequency; L is the inductance value. In this design, ΔI_L is very small, less than 5 *mA*, and ESR is less than 1 Ω, so this part of the power comsumption P_{ESR} is negligible.

Conduction Loss of the MOSFETs High side MOSFET causes conduction losses when the high side MOSFET turns on and the low side MOSFET turns off during DT:

$$P_{CONH} = I_{OUT}^2 \cdot R_{ONH} \cdot \frac{V_{OUT}}{V_{IN}} \tag{5.31}$$

When the low side MOSFET is turned off and the high side MOSFET is turned on during (1-D)T, the low side MOSFET causes conduction losses:

$$P_{CONL} = I_{OUT}^2 \cdot R_{ONL} \cdot \left(1 - \frac{V_{OUT}}{V_{IN}}\right) \tag{5.32}$$

where I_{OUT} is the output DC current; R_{ON} is the high side MOSFET on-resistance; R_{OFF} is the low side MOSFET on-resistance; V_{IN} is the input voltage; V_{OFF} is the output voltage.

The above two equations ignore the AC inductor current, as shown in the previous section, because the AC current is too small.

In this design, in order to achieve compact size and low on-resistance, the DMC2038LVT is selected as a MOSFET, including NMOS Q1 and PMOS Q2. The NMOS on-resistance is 35 $m\Omega$ $@V_{GS} = -2.5\ V, I = 1\ A$. The PMOS on-resistance is 80 $m\Omega$ $@V_{GS} = 2.5\ V, I = 1\ A$.

Based on the above analysis, the conductance of the MOSFET is:

$$P_{CON} = P_{CONH} + P_{CONL} = I_{OUT}^2 \cdot \left[\frac{V_{OUT}}{V_{IN}} \cdot (R_{ONH} - R_{ONL}) + R_{ONL} \right] \tag{5.33}$$

A lower input voltage results in higher MOSFET losses, and the minimum input voltage of 2.5 V is used as the input voltage in this design.

$$P_{CON} = P_{CONH} + P_{CONL} = 1\ A^2 \cdot \left[\frac{1.5}{2.5} \cdot (80\ m\Omega - 35\ m\Omega) + 35\ m\Omega \right] = 62\ mW. \tag{5.34}$$

There is one point worth noting here. In order to achieve low on-resistance, large ratio power transistors are often used, but this causes some problems. Larger MOSFETs can reduce the on-resistance and then reduce the conductance power loss, but increase the parasitic capacitance, which increases the capacitance power loss and vice versa. Therefore, this is a trade-off when choosing the MOSFET size.

Switching Loss of the MOSFETs When the power MOSFET is turned on and off, the switching losses of the MOSFET occur. Use a high-side MOSFET switch here, for example, as shown in Figure 5.18.

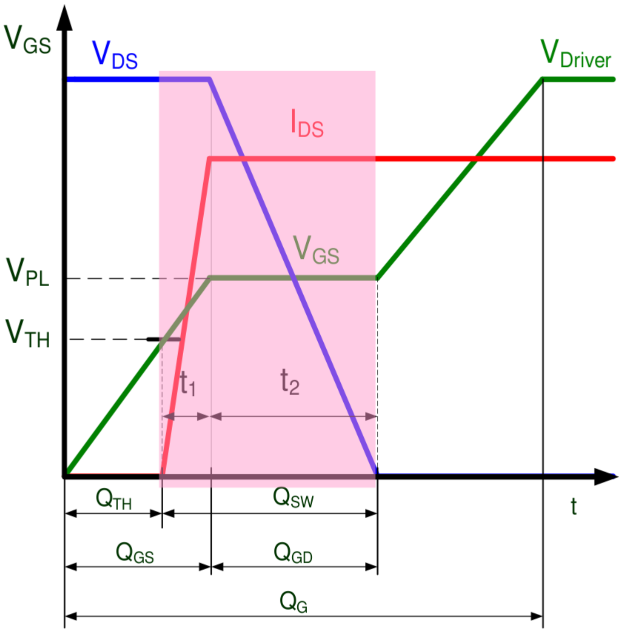


Fig. 5.18: High side MOSFET switching on process

where the green line is the gate voltage $\mathrm{V_{Driver}}$, the blue one is the drain and source voltage $\mathrm{V_{DS}}$, and the red one is the drain current $\mathrm{I_{DS}}$. During the gate voltage switching of the MOSFET, $\mathrm{V_{Driver}}$ and $\mathrm{V_{DS}}$ are not both zero in the $\mathrm{t_1}$ and $\mathrm{t_2}$ time, the pink area. In this case, the MOSFET will consume some power. The descending process is the same as the rising process. The high-side MOSFET switching loss is:

$$P_{SWH} = \frac{1}{2} \cdot V_{IN} \cdot I_{OUT} \cdot \left(t_{rH} + t_{fH}\right) \cdot f_{SW} \tag{5.35}$$

where V_{IN} is the input voltage; I_{OUT} is the output current; t_{rH} is the high side MOSFET rise time; t_{fH} is the fall time, and f_{SW} is the frequency. Under the same principle, the low side MOSFET power loss is:

$$P_{SWL} = \frac{1}{2} \cdot V_{D} \cdot I_{OUT} \cdot \left(t_{rH} + t_{fH}\right) \cdot f_{SW} \tag{5.36}$$

where V_D is the forward direction voltage of the low side MOSFET body diode. I_{OUT} and f_{SW} are the same with the high side MOSFET, and t_{rH} and t_{fH} are the rising time and falling time of the low side MOSFETs respectively. When the low side MOSFET turns on, the V_D is very low, so normally P_{SWL} is small and can be ignored. Therefore, the main power loss of the switch comes from the P_{SWH}.

From the data sheet of the DMC2038LVT, the rise time and fall time of Q2 PMOS are 12 *ns* and 13 *ns*, respectively. f_{SW} is 1 *MHz*, V_{IN} is 3.3 *V*, and I_{OUT} is 1 *A*. Substitute these parameters into the P_{SWH} equation. The power loss is 41 *mW*.

Reverse Recovery Loss of the Low Side MOSFET When the buck converter turns the high side MOSFET from OFF to ON, the body diode of the low side MOSFET reverses its current direction from the positive direction. This course will generate a reverse current as shown in Figure Ref Diode Power Loss.

In this figure, t_{rr} is the recovery time, and this power loss is:

$$P_{DIODE} = \frac{1}{2} \cdot V_{IN} \cdot I_{RR} \cdot t_{rr} \cdot f_{SW} \tag{5.37}$$

where V_{IN} is the input voltage; I_{RR} is the body diode recovery maximum current; f_{SW} is the switching frequency.

In this design, the datasheet does not provide I_{RR} and t_{rr}, but provides a reverse transfer capacitance of 65 *pF*, which may affect reverse current. Since the reverse

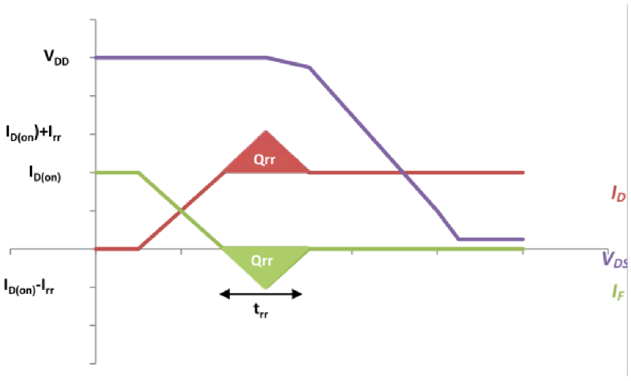


Fig. 5.19: Low side MOSFET body diode reverse recovery process

transfer capacitance is small, t_{rr} is only a few nanoseconds. Diode power loss is negligible.

Power Loss on the Parasitic Capacitance at the Drain Terminal When the current charges and discharges the capacitor, there is power loss even if there is no dissipative component. When charging a capacitor, the energy stored in the capacitor is:

$$P_{CAP} = \frac{1}{2} \cdot C \cdot V_{CAP}^2 \tag{5.38}$$

where C is the capacitance, V_{CAP} is the capacitor voltage.

During this charging process, there is a power loss that is the same as the power stored in the capacitor. Some of the dissipated power is in the form of heat sinking by series parasitic resistors, and other dissipated power is radiated through electromagnetic waves.

Whenever the high-side MOSFET is turned on, the input voltage charges the drain-side capacitor, so the power loss is:

$$P_{COSS} = \frac{1}{2} \cdot (C_{OSSL} + C_{OSSH}) \cdot V_{IN}^2 \cdot f_{SW} \tag{5.39}$$

where $C_{OSSL} = C_{DSL} + C_{GDL}$, and $C_{OSSH} = C_{DSH} + C_{GDH}$.
C_{OSSL} is the low side MOSFET output capacitance;
C_{DSL} is the low side MOSFET drain source capacitance;
C_{GDL} is the low side MOSFET gate drain capacitance;

C_{OSSH} is the high side MOSFET output capacitance;
C_{DSH} is the high side MOSFET drain source capacitance;
C_{GDH} is the high side MOSFET gate drain capacitance;
V_{IN} is the input voltage;
f_{SW} is the frequency.

C_{OSS} is a variable capacitor, depending on the drain source voltage. From the data sheet, C_{OSS} can be checked as shown in Figure 5.20.

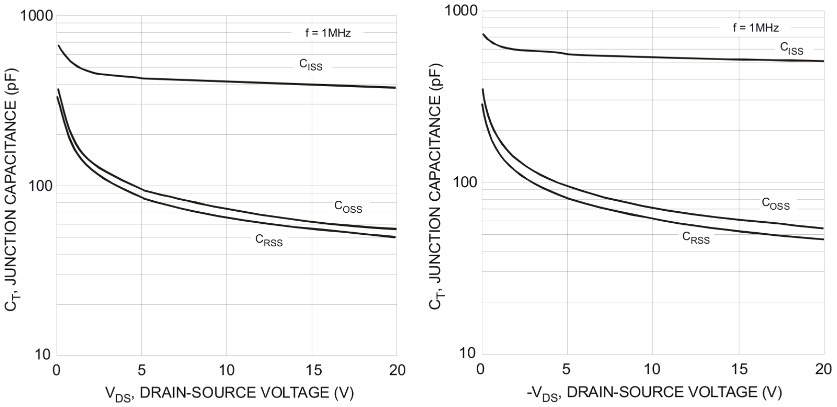


Fig. 5.20: DMC2038LVT Q1 NMOS parasitic capacitance (left) and Q2 PMOS parasitic capacitance (right)

It can be seen that when the drain source voltage is 2.5 V, C_{OSSH} is about 100 pF, and C_{OSSL} is about 120 pF. Therefore, when f_{SW} is 1 MHz, P_{COSS} is approximately 1 mW.

Power Loss on the Parasitic Capacitance at the Gate Terminal When the PWM driver turns the power transistors on and off, they charge and discharge the gate capacitances of the high-side and low-side MOSFETs as shown in Figure 5.21.

When these output capacitors are not charged, these charges will flow into the output, so the power stored in the output capacitors is not wasted. But when charging and discharging the gate capacitance, all of the power stored in the capacitor will be wasted. So the gate loss is:

$$P_{COSS} = (Q_{gH} + Q_{gL}) \cdot V_{gs} \cdot f_{SW} \tag{5.40}$$

$$P_{COSS} = (C_{GSL} + C_{GSH}) \cdot V_{gs}^2 \cdot f_{SW} \tag{5.41}$$

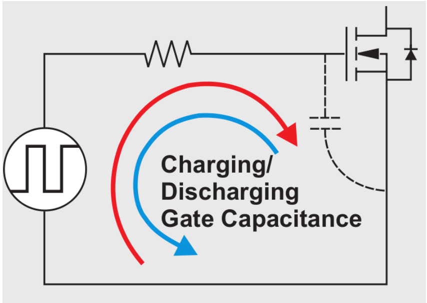


Fig. 5.21: Switch MOSFET gate power loss

Q_{gH} is the gate charge of the high side MOSFET; Q_{gL} is the gate charge of the low side MOSFET; C_{GSH} is the gate capacitance of the high side MOSFET; C_{GSL} is the gate capacitance of the low side MOSFET; V_{gs} is gate driver voltage. f_{SW} is the frequency.

C_{GS} is C_{ISS}-C_{RSS}. From Figure 5.21, C_{GSH} is about 500 pF @2.5 V, and C_{GSL} is about 400 pF @2.5 V. Therefore, when the frequency is 1 MHz, the gate loss is approximately 6 mW.

Power Loss Result from the Dead Time There is a disadvantage to using a low side MOSFET instead of a diode. When both MOSFETs are turned on at the same time, a short circuit occurs between the input voltage and ground. To avoid this, the system requires a dead time controller that can be inserted for a short period of time between the N-type MOSFET action and the P-type MOSFET action.

The low-side MOSFET has turned off for a short time (t_{Dr}) before the high side MOSFET turns on; the high side MOSFET has turned off for a short time (t_{Df}) before the low-side MOSFET turns on. During the dead time, both switching MOSFETs are turned off and the inductor current will flow through the low side MOSFET. The equation is:

$$P_{DEAD} = V_D \cdot I_{OUT} \cdot \left(t_{Dr} + t_{Df}\right) \cdot f_{SW} \tag{5.42}$$

where V_D is the forward direction voltage of low side MOSFET body diode; I_{OUT} is the output current; t_{Df} and t_{Dr} have been defined just now. In this design, the dead time t_{Df} and t_{Dr} are set to 30 ns, so the dead time power consumption is approximately 40 mW.

According to the above discussion, the DC-DC power dissipate is the sum of P_{IND} 190 mW, P_{ESR} 62 mW, P_{SW} 41 mW, P_{COSS} 1 mW,$P_{G}ATE$ 6 mW,P_{DEAD} 40 mW. The sum is 0.34 W. The output power is $1.5\ V \cdot 1\ A = 1.5\ W$, so when output current is 0.6 A, the efficiency is $1.5\ W/(1.5\ W + 0.34\ W) = 81.5\%$.

Dead Time Controller

The purpose of the dead time controller has been discussed in the last paragraph. In this paragraph, the detailed design will appear [43]. The dead time controller architecture is shown in Figure 5.22.

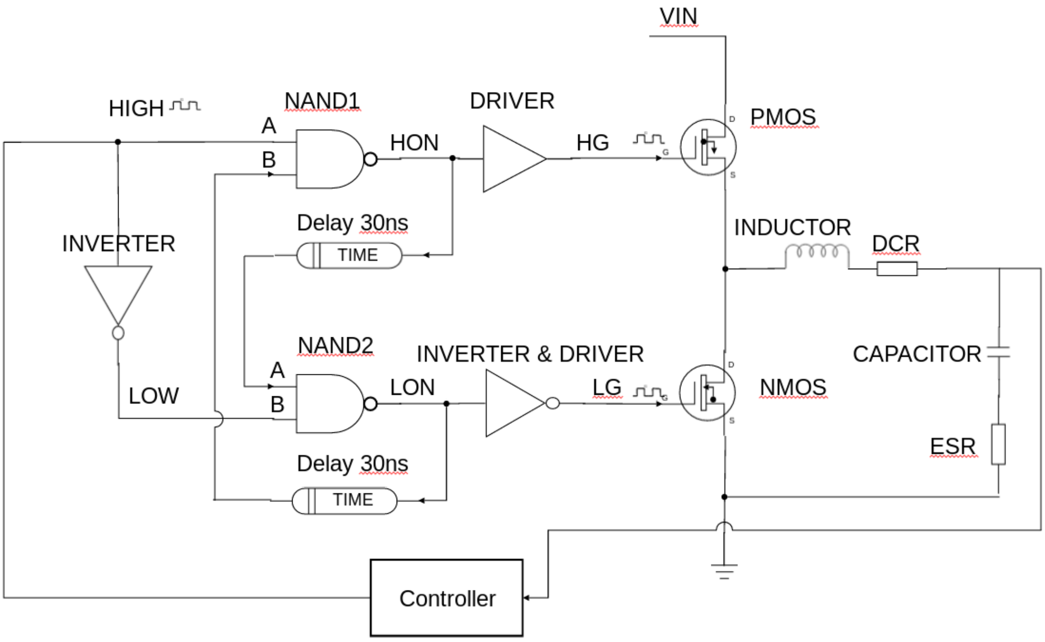


Fig. 5.22: Dead time controller architecture

The architecture uses NAND logic cells and delay cells to complete the control core. The NAND logic real table is shown in Table 5.2.

Then the power transistor state changes from the ON duty to the OFF duty, which means PMOS will change on to off, and NMOS changes from off to on. To avoid NMOS and PMOS are on-state at the same time, PMOS should switch off firstly, and then after a short time, the NMOS switch on. The course is as follows.

When the power MOSFET is in the ON state, this means that the PMOS is turned on and the NMOS is turned off. The HON signal is turned off and the LON signal is

Table 5.2: Truth table of the NAND logic

A	B	Y
0	0	1
0	1	1
1	0	1
1	1	0

turned on. The controller output signal HIGH is turned on and the LOW signal is turned off. Then, the power transistor state changes from the on-duty to the off-duty, which means that the PMOS will be turned on to off, and the NMOS will change from off to on. To prevent the NMOS and PMOS from being turned on at the same time, the PMOS should be turned off first, and then after a short time, the NMOS is turned on. The course is as follows.

Firstly, the controller output signal changes from on to off, which means HIGH is on, and LOW is turned on almost at the same time. At this time, the A terminal of NAND2 is 0, and the B terminal is 1, so the output LON does not change, and it is still 1. The A terminal of NAND1 is 0, and the B terminal is 1, so the output is in NAND1, and HON will change from 0 to 1. This can turn off the PMOS firstly.

After the delay of 30 *ns*, HON signal 1 is transmitted to NAND2, so the A and B terminals of NAND2 are both 1, the LON changes from 1 to 0. After the inversion, the NMOS gate LG signal changes from 0 to 1, which can be turn on, after 30 *ns* delay of the PMOS turns off.

This process is similar to the above action when the system state changes from OFF duty to ON duty. Therefore, the dead control unit can achieve a dead time of 30 *ns* .

The delay function is implemented by the following circuit, as shown in Figure 5.23. The delay function consists of four inverters in series and a low-pass filter. The four inverters do not change the phase of the input signal. However, it can amplify the drive current to make the signal clearer, thereby reducing the rise and fall times of the power transistor gate, and reducing the gate power loss.

The delay time is mainly from the RC low-pass filter, rarely from the inverter. The capacitor is located at the gate of the last inverter, so it takes some time to accumulate the charge on the capacitor. When the accumulated power reaches a

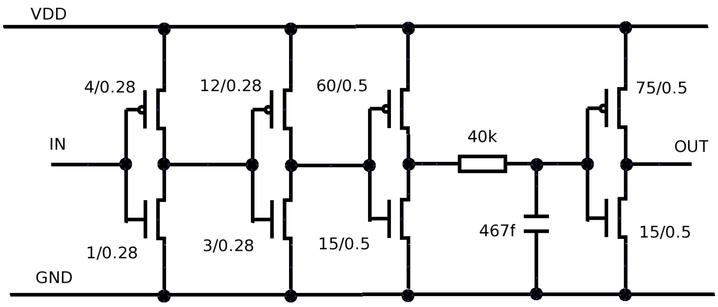


Fig. 5.23: Delay function schematic

certain point, the last inverter starts to work. The resistor and capacitor form an RC constant of approximately 19 *ns*.

Although inaccurate, the RC delay time, along with the four inverter times, is approximately 20 *ns*, which is almost identical to the schematic simulation results. After the layout, the post-simulation variation changes the delay time from 20 *ns* to 30 *ns*, due to the parasitic resistance and capacitance increasing the RC constant.

5.3.2 Transfer Function and the Feedback Compensation

In this design, the feedback method selects the voltage mode. The key to feedback design is the compensation circuit. Firstly, an open loop transfer function must be given. The compensation design is based on functionality. As mentioned earlier, the system is divided into three parts. The converter power stage is analyzed as follows.

Open-Loop Transfer Function

As previously mentioned, the system includes a converter power stage block, a modulator block and a compensator block. The open loop transfer function consists of two block transfer functions: a converter power stage block and a modulator block.

Converter Power Stage Transfer Function The block of the converter power stage is shown in Figure 5.24. To make it easier to calculate, the function can be divided into two parts. One is from point D to point SW, and the other is from point SW to V_{out} point.

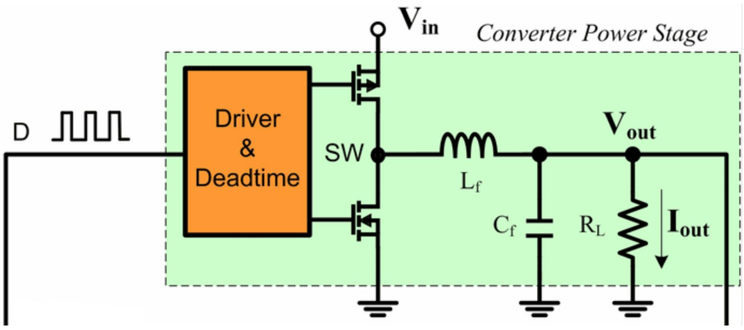


Fig. 5.24: Converter power stage block

The first part is very simple. The D signal is a square wave with a duty cycle of D. From the previous section the SW DC section can be gotten:

$$V_{SW} = D \cdot V_{IN} \tag{5.43}$$

$$\text{Gain}_{part1} = \frac{V_{SW}}{D} = V_{IN} \tag{5.44}$$

The second part of the function depends on the indcutor, capacitance and parasitic resistance. For more detailed functionality, the detailed mode is shown in Figure 5.25.

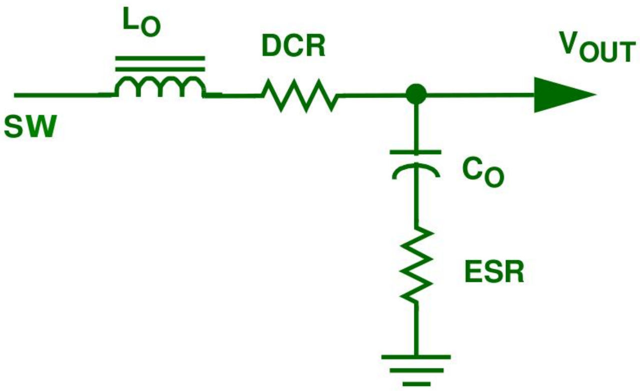


Fig. 5.25: The second part of the converter power stage block

The function of the second part is:

$$\text{Gain}_{part2} = \frac{1 + s \cdot \text{ESR} \cdot \text{C}_{\text{OUT}}}{1 + s \cdot (\text{ESR} + \text{DCR}) \cdot \text{C}_{\text{OUT}} + s^2 \cdot \text{L}_{\text{OUT}} \cdot \text{C}_{\text{OUT}}} \tag{5.45}$$

Merge the two parts together, and then the whole converter power stage block transfer function can be gotten:

$$\mathrm{Gain}_{power} = V_{IN} \cdot \frac{1 + s \cdot \mathrm{ESR} \cdot \mathrm{C_{OUT}}}{1 + s \cdot (\mathrm{ESR} + \mathrm{DCR}) \cdot \mathrm{C_{OUT}} + s^2 \cdot \mathrm{L_{OUT}} \cdot \mathrm{C_{OUT}}} \tag{5.46}$$

Modulator Block Transfer Function The modulator mode consists of a PWM comparator and a ramp sawtooth generator as shown in Figure 5.26.

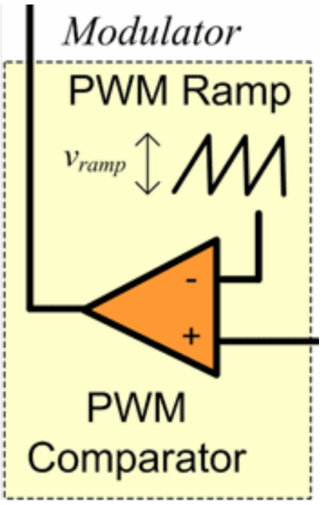


Fig. 5.26: Illustration of the modulator block

In this modulator, the error signal V_e from the error amplifier will be connected to the non-inverting terminal and the ramp sawtooth will be connected to the inverting terminal. The comparator function compares ramp sawtooth and V_e, and outputs a square with some duty cycle. It is shown in Figure 5.27.

According to the triangular equivalence principle, the duty cycle D is equal to the ratio of V_e over V_{osc}. V_{osc} is the maximum ramp sawtooth amplitude.

$$D = \frac{V_e}{V_{osc}} \tag{5.47}$$

The modulator block transfer function $Gain_{comp}$ from Ve to D is:

$$\mathrm{Gain}_{comp} = \frac{1}{V_{ramp}} \tag{5.48}$$

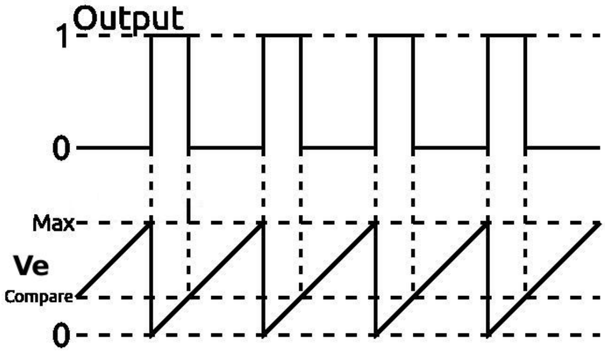


Fig. 5.27: Illustration of the PWM generator

Combining the equations (5.49) and (5.46), the open-loop transfer function Gain_{open} is:

$$\text{Gain}_{open} = \frac{1}{V_{osc}} \cdot V_{IN} \cdot \frac{1 + s \cdot \text{ESR} \cdot \text{C}_{\text{OUT}}}{1 + s \cdot (\text{ESR} + \text{DCR}) \cdot \text{C}_{\text{OUT}} + s^2 \cdot \text{L}_{\text{OUT}} \cdot \text{C}_{\text{OUT}}} \tag{5.49}$$

From Gain_{open}, it can be found that the open loop system has a double pole and a zero point. The double pole comes from the inductor and capacitor, which makes the system unstable. The zero comes from ESR and capacitor, which may improve the phase margin. The low frequency gain is:

$$\text{Gain}_{low} = \frac{V_{IN}}{V_{osc}} \tag{5.50}$$

By solving the numerator and denominator roots, it can get the frequency of the zero pole F_{ESR} and the double poles F_{LC} respectively:

$$F_{ESR} = \frac{1}{2 \cdot \pi \cdot ESR \cdot C_o} \tag{5.51}$$

$$F_{LC} = \frac{1}{2 \cdot \pi \cdot \sqrt{L_o \cdot C_o}} \tag{5.52}$$

The bode plot of the open-loop system is shown in Figure 5.28.

The low-frequency amplitude is V_{in}/V_{osc}. As the frequency increases, the magnitude will drop sharply with a slope of -40 dB/dec to the zero point frequency F_{ESR}.

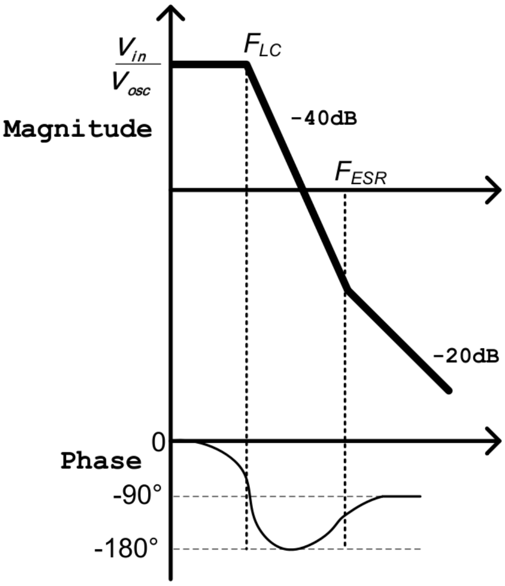


Fig. 5.28: Bode plot of the open-loop system

The left zero point will cancel one pole, so the slope becomes -20 dB/dec. As the amplitude decreases, the phase also drops rapidly. At F_{LC}, the phase will drop 90°.

Zero point F_{ESR} can make the phase better. But F_{ESR} depends on capacitance and ESR. ESR is usually small, which can result in a high zero. So F_{ESR} is usually beyond the unity gain point. In this case, the zero point is not useful for improving the phase margin.

Due to the above analysis, the phase margin is always lower than 45°, and the open loop system will be unstable. It requires a compensation network to stabilize the system.

Compensation Network

The ideal compensation network should have the following conditions:

- When the loop gain magnitude dropping through the 0 dB (the magnitude is one), the slope should be 20 dB/dec;
- The phase margin should be larger than the 45°;
- The F_0 (the crossover frequency) as large as possible.

The required Bode plot is shown in Figure 5.29.

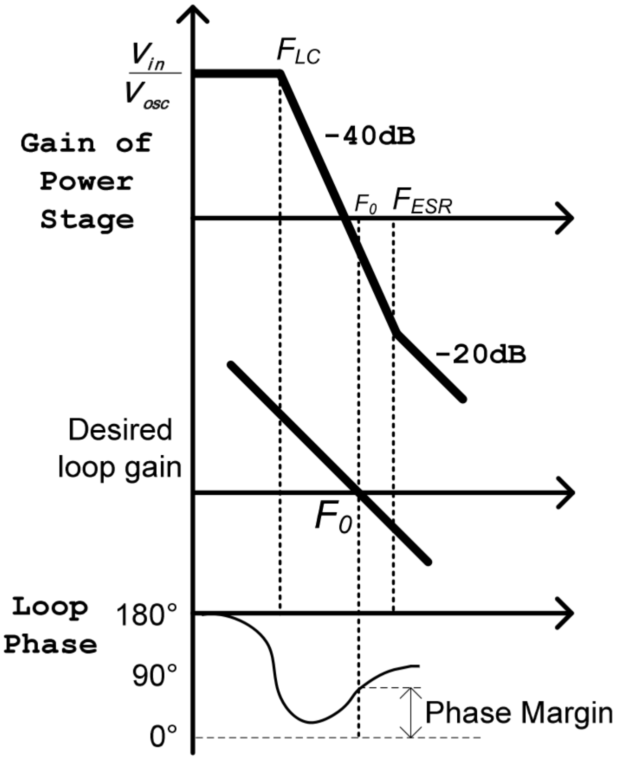


Fig. 5.29: Bode plot of the desired loop gain and phase

The first step in designing a compensation network is to fix the crossover frequency F_0, which is also the bandwidth of the closed loop or the bandwidth of the system. F_0 determines the system response speed.

There is a trade off when fixing F_0. If F_0 is set higher, the system will have a lower output impedance, and therefore respond faster for the output current and input voltage variation. But F_0 has an upper limit, which is half the switching frequency. According to the Nyquist sampling theory, the sampling frequency must be larger than twice the signal frequency. In this buck converter system, the ramp saw can be seen as the signal acquisition circuit and the error amplifier output V_e is treated as a sampled signal. When the system bandwidth is equal to or larger than half of the switching frequency, the ramp saw will not sample the useful signal. Some system signals are lost due to the higher frequency.

There are other limitations on the upper frequency limit. One of them is the slope matching principle, as shown in the figure ref match slope.

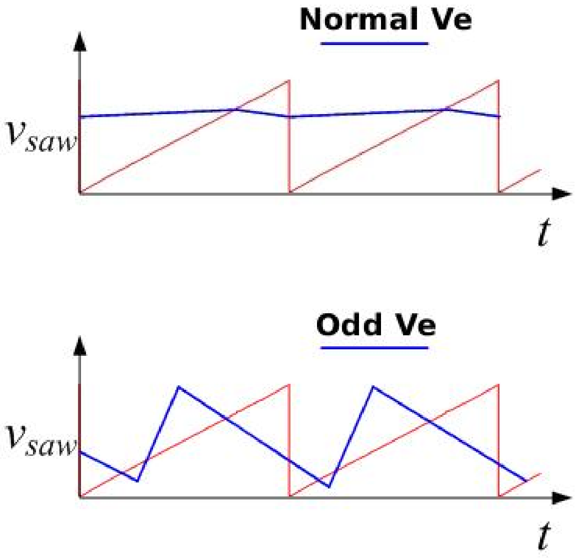


Fig. 5.30: Matching slope (up) and the no matching slope (down) voltage error

Bandwidth also suppresses high frequency ripple. When the bandwidth goes high, the error amplifier will output a high frequency signal, and then an odd number of *Ve* compared to the ramp saw will cause system instability. This is another reason why the bandwidth is set from 0.1 to 0.5 times the switching frequency.

Another important issue with bandwidth is the accuracy of the state space mean method. This method is the basis of small signal analysis. Use this method to analyze the PWM generator as shown in Figure ref state space.

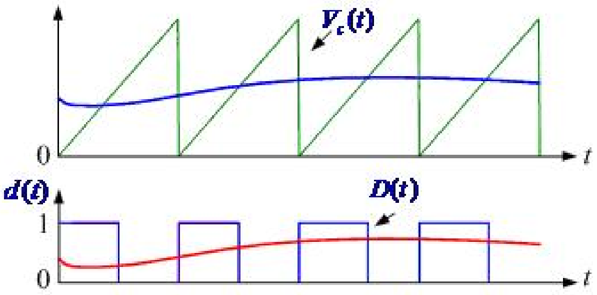


Fig. 5.31: State space mean method disposing the PWM

When the switching frequency is 100 kHz, and F_{LS} is 10 kHz, the corresponding duty cycle can be gotten by the PWM module. Using the state space mean method, the duty cycle is a 10 kHz signal, which is the red line in Figure 5.31. So the PWM can be considered as a ratio part. This has been discussed in the calculation of the transfer function of a PWM modulator. But it is not exactly the case. When doing the Fourier analysis to the duty cycle, the duty cycle not only includes the 10 *kHz* wave but also includes the 90 *kHz*, 110 *kHz*, 190 *kHz*, ... and so on.

The accuracy of the state space means method (or small signal analysis) depends on the suppression degree of the non-fundamental components. It is clear that the lower bandwidth can produce stronger suppression of non-fundamental components, and then the state space means method become more accuracy. When the bandwidth beyond the 1/5 of the switching frequency, the mode of the stage space mean will have the large difference with the actual situation.

Based on the above analysis, the bandwidth is set from 0.1 to 0.2 times the switching frequency. The next step is to choose the compensation network type. There are two types of compensation networks commonly used for buck converter designs: Type II and Type III. Which compensation to choose depends on the output capacitor and its ESR. The ESR and capacitor form the left zero.

$$F_z = -\frac{1}{C_{out} \cdot R_{ESR}} \tag{5.53}$$

Zero can improve phase margin and amplitude. Different kinds of capacitors have different ESRs, so the zero point will have a wide range. Table 5.3 shows different compensation types. Depending on the capacitor, choose a corresponding compensation type.

If the output capacitor is designed on-chip, type III compensation should be used. Because both MIM caps and MOM caps have low ESR, they are similar to ceramic capacitors.

Table 5.3: Different compensation types according to the ESR zero position

Compensation type	The ESR zero point place	Typical output capacitor
Type II(PI)	$F_{LC} < F_{ESR} < F_0 < F_s/2$	Electrolytic, POS-Cap, SP-Cap
Type III-A (PID)	$F_{LC} < F_0 < F_{ESR} < F_s/2$	POS-Cap, SP-Cap
Type III-B (PID)	$F_{LC} < F_0 < F_s/2 < F_{ESR}$	Ceramic

Type II Compensation Network A good compensation system should have high gain at low frequencies and low gain at high frequencies. The high gain at low frequencies reduces the DC error between the output voltage and the designed voltage. Low gain at high frequencies reduces high frequency noise. In addition to this, the high frequency gain of the amplifier is also limited. The compensation schematic is shown in Figure 5.32.

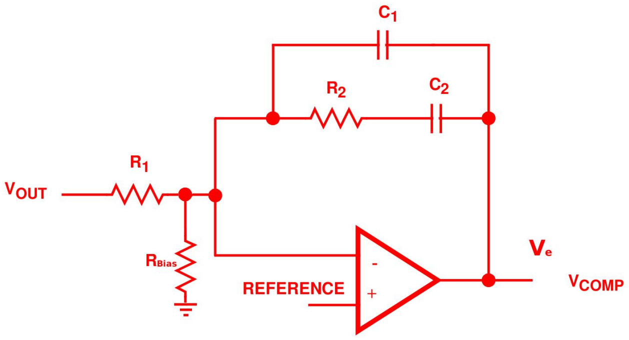


Fig. 5.32: Schematic of the compensation type II

The transfer function of the type II compensation $H(s)$ is:

In open-loop gain, the low-frequency gain is V_{IN}/V_{osc}, which is not enough to make a small difference between the output voltage and the desired output voltage. A common method of enhancing low frequency gain is the integrator. The integrator can be implemented by a capacitor between the inverting terminal and the output terminal of the error amplifier. The integrator forms a pole at the origin. The transfer function of type II compensation $H(s)$ is:

$$H(s) = \frac{V_e}{V_{OUT}}(s) = -\frac{1 + R_2 \cdot C_2 \cdot s}{R_1 \cdot s \cdot (C_2 \cdot C1) \cdot \left(R_2 \cdot \frac{C_2 \cdot C_1}{C_2 + C_1} \cdot s + 1\right)} \tag{5.54}$$

To get a clearer pole and zero, this equation can be rewritten as:

$$H(s) = -\frac{1}{R_1 \cdot C_1} \cdot \frac{\left(s + \frac{1}{R_2 \cdot C_2}\right)}{s \cdot \left(s + \frac{C_1 + C_2}{R_2 \cdot C_1 \cdot C_2}\right)} \tag{5.55}$$

This equation has two left poles and one left zero point. they are:

$$F_{p1} = 0 \tag{5.56}$$

$$F_{p2} = -\frac{C_1 + C_2}{R_2 \cdot C_1 \cdot C_2} \tag{5.57}$$

$$F_{z1} = -\frac{1}{R_2 \cdot C_2} \tag{5.58}$$

The amplitude and phase variation is shown in Figure 5.33.

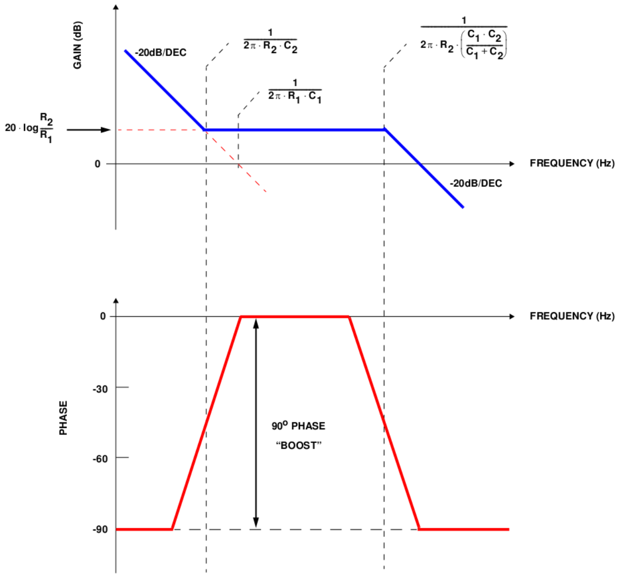


Fig. 5.33: Amplitude and phase of the type II compensation

The first pole is at the origin and is introduced by the integrator. Although the origin can greatly improve the low frequency gain, it also has a disadvantage. It will move the phase -90° for all frequencies. Therefore, the compensation network needs to provide a left zero to compensate for the loss of phase.

If the ESR zero point is high, it can not cancel one of the double poles. At the same time, the zero should not be too low, otherwise the cross-section frequency will be close to the second pole, which will also reduce the phase margin. So the zero point must be set to suitable position for a larger phase margin.

Typically, the zero point is set from 0.1 to 0.75 times the resonant frequency F_{LC}. Due to the zero point, the amplitude does not drop continuously with the slope of 20 *dB*, and the phase will rise by +90°.

Then the double pole caused by the *LC* resonant frequency will work, which causes the amplitude to drop sharply because the slope is -40 *dB* and the phase shifts to -180°. This is a bad situation. The double pole can be compensated by the ESR zero point. Type II compensation is used when the ESR zero is at a relatively low frequency. The ESR zero will cause the system to increase the slope of 20 *dB* and increase the phase of 90°.

Due to the ESR zero point, the amplitude will drop cross the unit line as the -20 *dB* slope with an acceptable phase margin. After that, the second pole will work. In order to reduce the second pole effect, the second pole should be further away from the crossover frequency.

The second pole has another upper limit. The limit is the pole should not exceed the switching frequency F_{SW}. The second pole plays another role to reduce the high-frequency noise, especially the switch frequency noise. The second pole should be kept at a certain distance from the switching frequency. As a compromise, half of the switching frequency is used as the second pole.

The second pole can be expressed as:

$$F_{p2} = -\frac{C_1 + C_2}{R_2 \cdot C_1 \cdot C_2} = -\frac{\frac{C_1}{C_2} + 1}{R_2 \cdot C_1} \tag{5.59}$$

Compare the zero point with the second pole. Because of $F_{p2} \gg F_{z1}$ and $C_2 \gg C_1$, the second pole can be simplified:

$$F_{p2} \approx -\frac{1}{R_2 \cdot C_1} \tag{5.60}$$

When the error amplifier is ideal, the low frequency gain is infinite, but the actual error amplifier has limited gain and bandwidth. Therefore, the low frequency gain is limited by the error amplifier gain and is about 7 *dB* lower than its gain. The actual error amplifier limits the loop gain, so the DC component of the output voltage deviates from the setting value. The output impedance, which can influence the load regulation, cannot be further reduced.

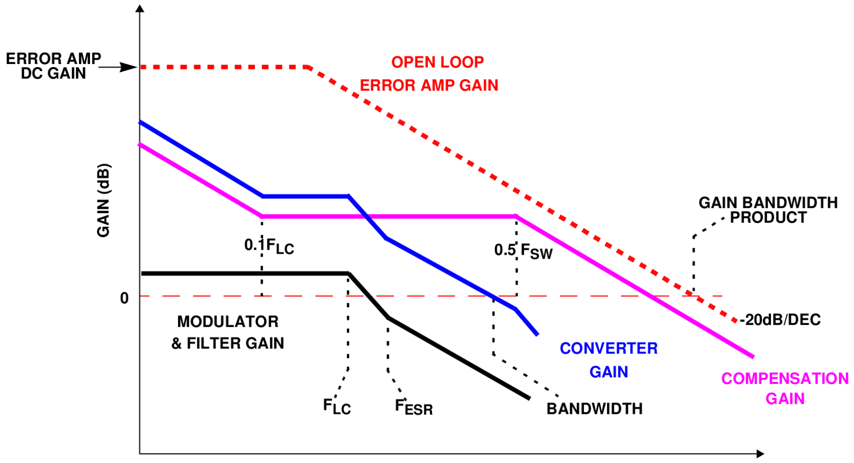


Fig. 5.34: Compensated system amplitude bode plot using the type II compensation

The compensated system amplitude Bode plot can be shown in Figure 5.34. The pink line is the compensation amplitude, the black one is the open-loop amplitude, and the blue one is the compensated amplitude which crosses over the unit line as -20 *dB* slope. The red dotted line is the open-loop error amplifier. As can be seen from the Bode plot, system characters are limited by the error amplifier function. Better performance can be designed when the error amplifier has a higher DC gain and a larger bandwidth product.

The equation also shows that the resistor R_{Bias} does not appear in the transfer function, that means R_{Bias} does not contribute to the loop gain. When V_{OUT} is equals to the reference voltage, R_{Bias} can be removed. If V_{OUT} is larger than the reference voltage, R_{Bias} is essential to set the output voltage. V_{OUT} can be obtained according to the following equation:

$$V_{OUT} = V_{Ref} \cdot \frac{R_1 + R_{Bias}}{R_{Bias}} \tag{5.61}$$

$$R_{Bias} = \frac{R_1 \cdot V_{Ref}}{V_{OUT} - V_{Ref}} \tag{5.62}$$

The whole closed-loop system is shown in Figure 5.35.

The closed-loop system transfer function is [44]:

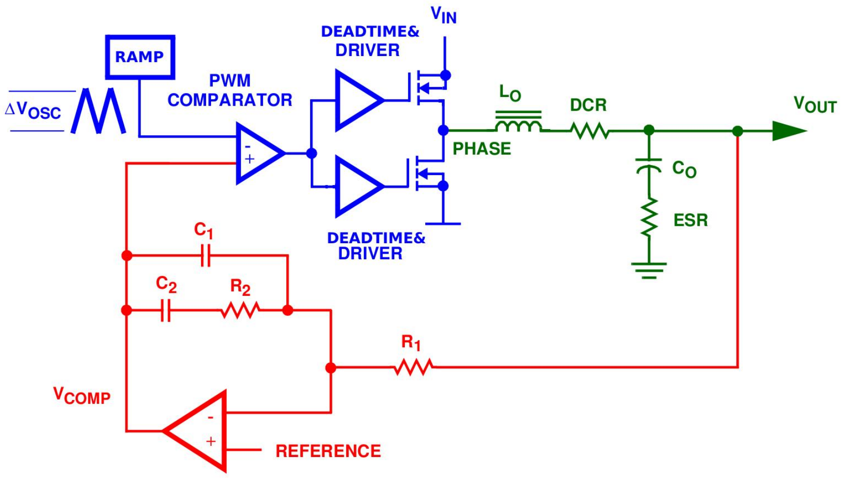


Fig. 5.35: Whole closed-loop schematic using the type II compensation

$$\mathrm{GAIN_{LOOP}} = \frac{1}{R_1 \cdot C_1} \cdot \frac{\left(s + \frac{1}{R_2 \cdot C_2}\right)}{s \cdot \left(s + \frac{C_1 + C_2}{R_2 \cdot C_1 \cdot C_2}\right)} \cdot \frac{V_{IN}}{\Delta V_{OSC}} \cdot \frac{1 + s \cdot ESR \cdot C_{OUT}}{1 + s \cdot (ESR + DRC) \cdot C_{OUT} + S^2 \cdot L_{OUT} \cdot C_{OUT}} \tag{5.63}$$

The different color portions of the closed loop equation correspond to the different color portions of the schematic. R_{Bias} does not appear here because this resistor does not contribute to loop gain. From the pole and zero analysis, it can be found that the key point of using Type II compensation is the ESR zero. The ESR zero should compensate for one of the two poles.

Type III Compensation Network Type III compensation is suitable for smaller ESR values. There are two types of type III compensation: Type III-A and Type III-B. The difference is the ESR zero point position. Both subtypes have the same schematic and only the resistor and capacitor values are different. The type III schematic is shown in Figure 5.36.

The transfer function of the type III compensation is:

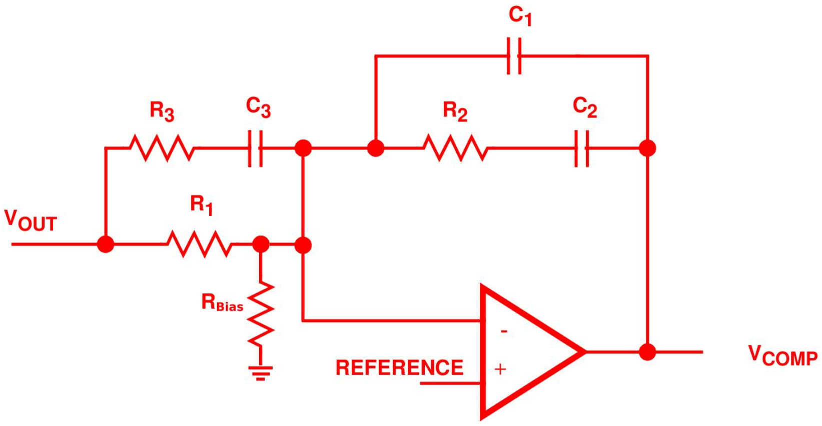


Fig. 5.36: Type III compensation schematic

$$H(s) = -\frac{R_1 + R_3}{R_1 \cdot R_3 \cdot C_1} \cdot \frac{\left(s + \frac{1}{R_2 \cdot C_2}\right) \cdot \left(s + \frac{1}{(R_1 + R_3) \cdot C_3}\right)}{s \cdot \left(s + \frac{C_1 + C_2}{R_2 \cdot C_1 \cdot C_2}\right) \cdot \left(s + \frac{1}{R_3 \cdot C_3}\right)} \tag{5.64}$$

Two zeros and three poles are given as below:

$$P_1 = 0; P_2 = -\frac{C_1 + C_2}{R_2 \cdot C_1 \cdot C_2}; P_3 = -\frac{1}{R_3 \cdot C_3} \tag{5.65}$$

$$Z_1 = -\frac{1}{R_2 \cdot C_2}; Z_2 = -\frac{1}{(R_1 + R_3) \cdot C_3} \tag{5.66}$$

The amplitude and phase Bode plot is shown in Figure 5.37.

First, analyze type III-A compensation. Like type II, the III-A origin pole caused by C_1 is used to increase the low frequency gain, which also reduces the phase margin by 90°. The first zero compensates for the origin pole and rolls back the phase margin of 90°. As with type II, the first zero choice has the same trade-off. A higher first zero will make the system more stable and slower, and vice versa. Considering the stability and speed, the first zero point is selected from 0.1 to 0.75 times the resonant frequency.

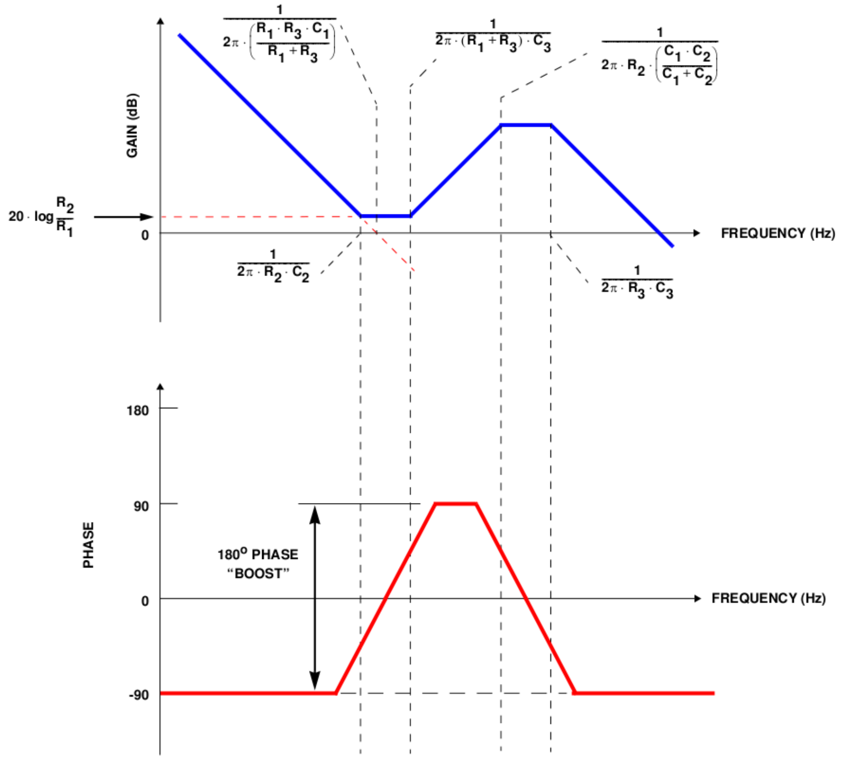


Fig. 5.37: Type III compensation bode plot

The purpose of the second zero is to compensate for the double pole of the open-loop function. The second zero causes the loop gain drop by -20 *dB*. The second pole is used to cancel the ESR zero, in order to make the gain drop continuously by -20 *dB* to unity gain. Beyond the unity gain, set the third pole to half the switching frequency to reduce the switching frequency noise. The pole and zero distributions in the open-loop and closed-loop functions are shown in Figure 5.38.

R_{Bias} also does not contribute to the loop gain. The whole closed-loop system of type III-A is shown in Figure 5.39. Different colors represent different parts.

The corresponding closed-loop transfer function is:

$$\mathrm{GAIN_{LOOP}} = \frac{R_1+R_3}{R_1\cdot R_3\cdot C_1}\cdot\frac{\left(s+\frac{1}{R_2\cdot C_2}\right)\cdot\left(s+\frac{1}{(R_1+R_3)\cdot C_3}\right)}{s\cdot\left(s+\frac{C_1+C_2}{R_2\cdot C_1\cdot C_2}\right)\cdot\left(s+\frac{1}{R_3\cdot C_3}\right)}\cdot\frac{V_{IN}}{\Delta V_{OSC}} \cdot\frac{1+s\cdot ESR\cdot C_{OUT}}{1+s\cdot(ESR+DRC)\cdot C_{OUT}+S^2\cdot L_{OUT}\cdot C_{OUT}} \tag{5.67}$$

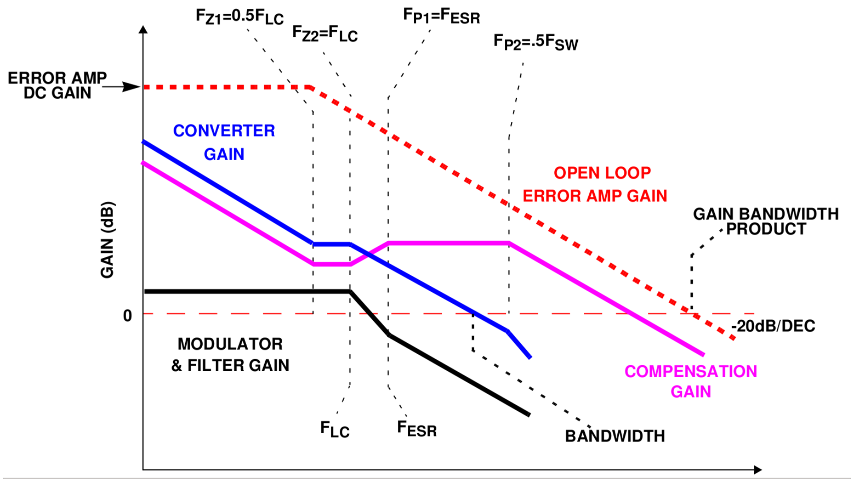


Fig. 5.38: Compensated system amplitude bode plot using the type III-A and the poles zeros distribution

The difference between types III-A and III-B is the ESR zero position. When ESR zero is larger than half of the switching frequency, it is useless to cancel the ESR zero point because it exceeds the unit frequency. So this situation will use one pole less. At this time, type III-B will be adopted.

The first original pole acts the same as the other compensation structures, which increases the low frequency gain. The first zero is also used to cancel the original pole. The role of the second zero is different from the type III-A, and it does not directly cancel one of the double zeros. It will work with the second pole to get a better phase margin. The last pole is the same as the III-A compensation, which is about half the switching frequency.

The second pole and second zero will constitute the lead compensator [45, 46]. Set the crossover frequency (closed-loop unit gain frequency) to F_0, and set the desired phase margin to θ, with:

$$F_{z2} = F_0 \cdot \sqrt{\frac{1-\sin\theta}{1+\sin\theta}} \tag{5.68}$$

$$F_{p2} = F_0 \cdot \sqrt{\frac{1+\sin\theta}{1-\sin\theta}} \tag{5.69}$$

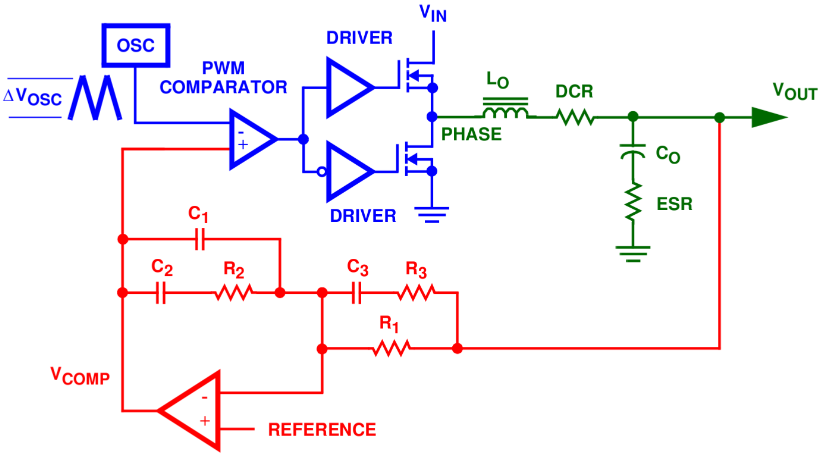


Fig. 5.39: Whole closed-loop schematic using the type III-A compensation

Usually the θ is set to 70°, which is the large practical phase-lead effect obtainable from the lead compensator. F_{z1} is usually set to half of the F_{z2}. So the third pole is set to half of the switching frequency.

This design is intended to use ceramic capacitors as output capacitors, which are not only small, but also have low ESR resistance. A lower ESR resistor reduces the voltage ripple. Type III is more suitable for this design. The feedback resistor and capacitor can be calculated based on the current parameters selected which is as follows.

$$\text{Input} = \begin{cases} V_{in} = 3.3\ V; \\ Vout = 1.5\ V; \\ Vref = 0.6\ V; \\ V_{ramp,osc} = 1.1\ V; \\ L_o = 68\ \mu H; \\ C_o = 1\ \mu F; \\ ESR(C_o) = 200\ m\Omega; \\ F_s = 1\ MHz; \\ I_{max} = 1\ A; \end{cases} \tag{5.70}$$

Firstly, calculate the ESR zero frequency and the *LC* double pole resonant frequency:

$$F_{LC} = \frac{1}{2\pi \cdot \sqrt{L_o \cdot C_o}} = \frac{1}{2 \cdot 3.14 \cdot \sqrt{68\ \mu H \cdot 1\ \mu F}} = 19.3\ kHz \tag{5.71}$$

$$F_{ESR} = \frac{1}{2\pi \cdot R_{ESR} \cdot C_o} = \frac{1}{2 \cdot 3.14 \cdot 200\ m\Omega \cdot 1\ \mu F} = 795\ kHz \tag{5.72}$$

F_o is from 1/10 to 1/5 of the switching frequency. In this design it is set to 1/10 of switching frequency, which is 100 kHz. According to Table 5.3, the frequency should satisfy this relationship:

$$F_{LC} < F_o < F_s/2 < F_{ESR} \tag{5.73}$$

Therefore, type III-B is more suitable for this design. When the phase margin is set to approximately 70°, the second pole and zero frequency are:

$$F_{z2} = F_o \cdot \sqrt{\frac{1 - \sin 70^\circ}{1 + \sin 70^\circ}} = 167\ kHz \cdot 0.176 = 29.4\ kHz \tag{5.74}$$

$$F_{p2} = F_o \cdot \sqrt{\frac{1 + \sin 70^\circ}{1 - \sin 70^\circ}} = 167\ kHz \cdot 5.67 = 946\ kHz \tag{5.75}$$

$$F_{z1} = \frac{1}{2} \cdot F_{z2} = \frac{1}{2} \cdot 29.4\ kHz = 14.7\ kHz \tag{5.76}$$

$$F_{p3} = \frac{1}{2} \cdot F_s = \frac{1}{2} \cdot 1\ MHz = 500\ kHz \tag{5.77}$$

Then the specific parameters can be calculated. There is a free choice for C_3, which is inversely proportioned to the resistance. In silicon designs, capacitor typically require a larger area than resistors. Therefore, when they are inversely proportion, it is better to choose a smaller capacitor and a larger resistor. This design chooses the $C_3 = 20\ pF$. R_3 can be obtained as follows:

$$R_3 = \frac{1}{2\pi \cdot C_3 \cdot F_{p2}} = \frac{1}{2 \cdot 3.14 \cdot 20\ pF \cdot 946\ kHz} = 8.4\ k\Omega \tag{5.78}$$

The next step is to calculate R_1. According to the second zero, there is:

$$R_1 + R_3 = \frac{1}{2\pi \cdot C_3 \cdot F_{z2}} = \frac{1}{2 \cdot 3.14 \cdot 20\ pF \cdot 29.4\ kHz} = 270.9\ k\Omega \tag{5.79}$$

$$R_1 = 270.9\ k\Omega - 8.4\ k\Omega = 262.5\ k\Omega \tag{5.80}$$

R_{Bias} is equal to $R_3 \cdot 2/3$ in order to make the output twice the reference voltage. At the loop unit gain frequency, the two zeros are lower than F_0, and ω will play a major roles. And the two poles are larger than F_0, the constants will play a major roles. So the loop gain can be simplified:

$$1 \approx \frac{R_1 + R_3}{R_1 \cdot R_3 \cdot C_1} \cdot \frac{s_0 \cdot s_0}{s_0 \cdot \frac{C_1 + C_2}{R_2 \cdot C_1 \cdot C_2} \cdot \frac{1}{R_3 \cdot C_3}} \cdot \frac{V_{IN}}{\Delta V_{OSC}} \cdot \frac{1}{s_0^2 \cdot L_o \cdot C_o} \tag{5.81}$$

When $R1$ is great larger than R_3, and C_2 is great larger than C_1, there is:

$$R_2 \cdot C_3 = 2\pi \cdot F_0 \cdot C_o \cdot L_0 \cdot \frac{\Delta V_{OSC}}{V_{IN}} \tag{5.82}$$

So to make F_o at desired position, R_2 is:

$$\begin{aligned} R_2 &= 2\pi \cdot F_0 \cdot C_o \cdot L_0 \cdot \frac{\Delta V_{OSC}}{V_{IN}} \cdot \frac{1}{C_3} \\ &= 2 \cdot 3.14 \cdot 100\ kHz \cdot 1\ \mu F \cdot 68\ \mu H \cdot \frac{1.1\ V}{3.3\ V} \cdot \frac{1}{20\ pF} = 712\ k\Omega \end{aligned} \tag{5.83}$$

R_2 and C_2 makes up the first zero, so there is:

$$C_2 = \frac{1}{2\pi \cdot R_2 \cdot F_{z1}} = \frac{1}{2 \cdot 3.14 \cdot 712\ k\Omega \cdot 14.7\ kHz} = 15.2\ pF \tag{5.84}$$

According to the last pole at 500 kHz, there is:

$$\frac{C_1 \cdot C_2}{C_1 + C_2} = \frac{1}{2\pi \cdot F_{p3} \cdot R_2} = \frac{1}{2 \cdot 3.14 \cdot 712\ k\Omega} = 0.447\ pF \tag{5.85}$$

$$C1 = \frac{C_2 \cdot 0.447\ pF}{C_2 - 0.447\ pF} = \frac{15.2\ pF \cdot 0.447\ pF}{15.2\ pF - 0.447\ pF} = 0.46\ pF \tag{5.86}$$

So the following parameters can be gotten:

$$\text{Output} = \begin{cases} R_1 = 262.5\ k\Omega; \\ R_2 = 712\ k\Omega; \\ R_3 = 8.4\ k\Omega; \\ C_1 = 0.46\ pF; \\ C_2 = 15.2\ pF; \\ C_3 = 20\ pF; \end{cases} \tag{5.87}$$

Up to now, all key parameters have been calculated. Although the modulator part, including the PWM comparator and the Ramp generator, has some complicated circuits, it does not have complex calculation parameters. So this part will be discussed in the next section.

5.3.3 Radiation-Aware Design

Similar to LDO and bandgap, the buck DC-DC radiation effects mainly comes from the TID and SEL. Therefore, the radiation-aware methods mainly focuse on these aspects.

In order to reduce the leakage current caused by the TID, most NMOS transistors use guardrings. Due to the relatively thin thickness of polysilicon, the 65 nm CMOS technology allows the circuit to withstand up to 100 *krad*(Si) TID, which is suitable for most radiation applications, including BESIII experiments.

For SEL radiation, most of the spare space is filled by the substrate connector and the nwell connector when laying out the layout. Another method is to properly increase the transistor distance.

In this paper, the power transistors are off-chip. A complete simulation analysis could not be completed. Based on the above design and analysis, the buck DC-DC can work normally in the BESIII experiment.

5.4 Simulation Results

In the last section, some key parameters are designed. At the same time, this section will simulate the different parts of the buck converter. There are several circuits that can be simulated. One is PWM generator simulation, one is dead time simulation, the other is error amplifier and compensation circuit simulation, and the third is the whole buck converter system simulation.

5.4.1 Simulation of the Modulator Architecture

The PWM generator consists of a comparator and a ramp generator. The comparator should have a fast reflection speed and the comparator consists of a simple 7-transistor amplifier and 4 inverters. It is shown in Figure 5.40. Since the power supply is 1.5 V, some interface transistors use 2.5 V I/O transistors. The schematic and layout of the comparator are shown in Figure 5.40. The length and width are 43 μm and 15 μm, respectively.

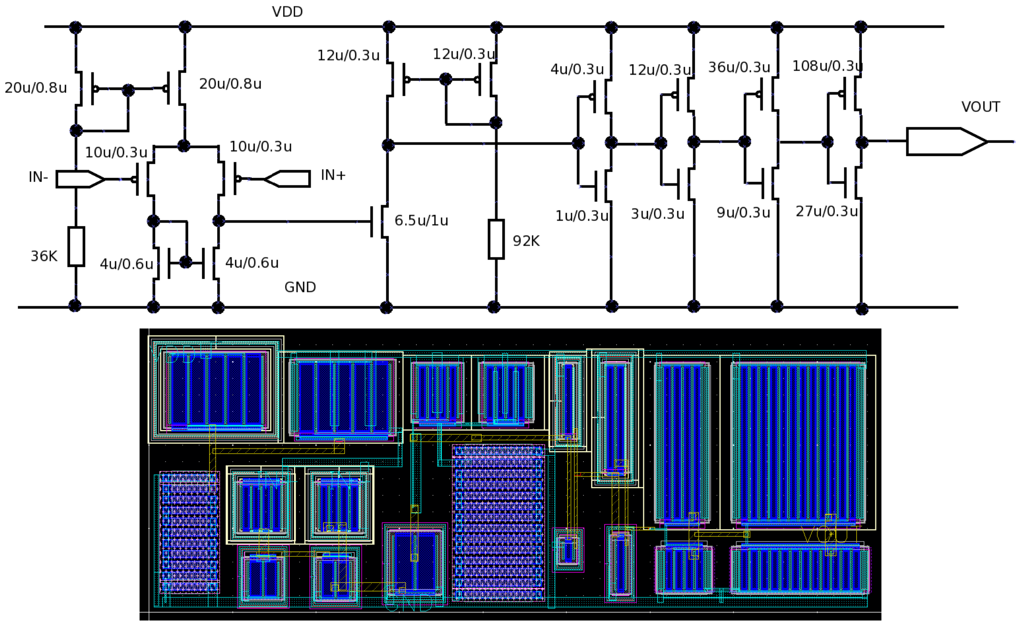


Fig. 5.40: Comparator schematic and layout

The four inverters have two roles. One is to make the comparator have more current drivers, and the other is to make the output voltage change abruptly. When the transistors have less switching time, power consumption can be reduced. When only one inverter is used, it will cause the OPA to output a large capacitance, which will make the system unstable. So here are four cascaded inverters, the latter being larger than the previous one.

Another disadvantage of multi-inverters is that they delay the comparator. The comparator delay time consists of three parts: the OPA bandwidth, the OPA slew rate, and the inverter *RC* delay. In this design, the main part is the inverter *RC* delay, and the total delay time is about 4 *ns*, as shown in Figure 5.41. The yellow line is the input voltage and the red dotted line is the output voltage.

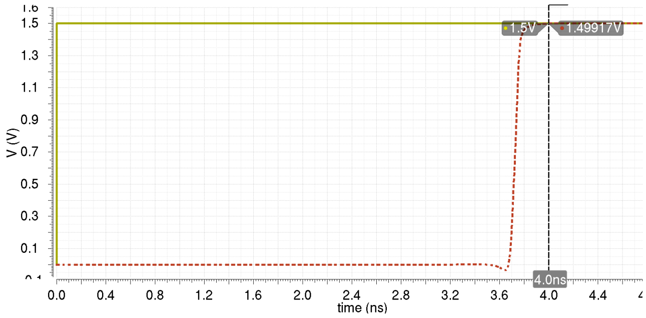


Fig. 5.41: Comparator delay time

Compared to the cycle time 1 μs, the 4 *ns* delay can be ignored.

Another important part of the PWM generator is the ramp generator, which uses a small current to charge the capacitor for the requested slope. The ramp generator architecture and layout are shown in Figure 5.42 and Figure 5.43, respectively. The layout is 155 μm length and 82 μm width.

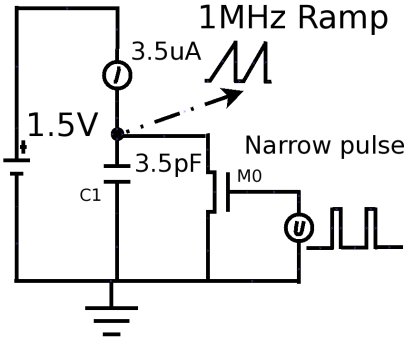


Fig. 5.42: Architecture of the ramp

The 3.5 μA current source charges the 3.18 pF capacitor, and the narrow pulse controls the switch NMOS M_0. Every 1 μs the narrow pulse resets M_1, which

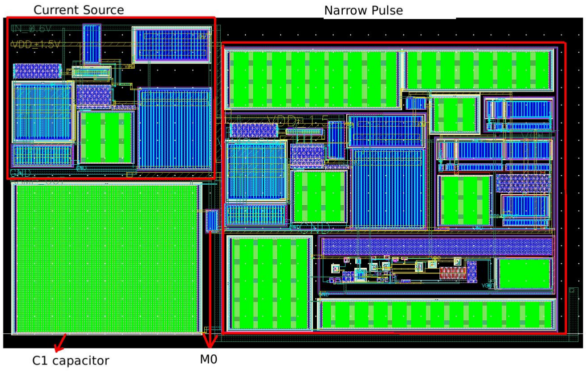


Fig. 5.43: Layout of the ramp

discharge the capacitor in a very short time. The current source architecture is the same as that used in the previous LDO chapter. The difference is in changing the output transistor ratio. There is one point should be noticed. The output PMOS has a very short time to connect to ground after discharging the capacitor, so the output pair transistors should use the 2.5 *V* transistor for safety.

Another key part is the narrow pulse generator, and the schematic is shown in Figure 5.44.

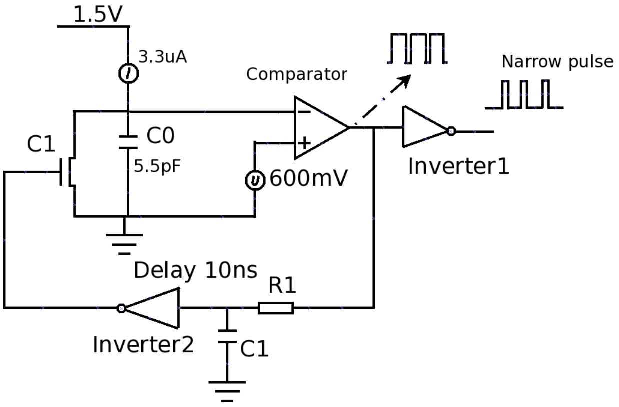


Fig. 5.44: Narrow pulse generator schematic

When the capacitor voltage is below the reference voltage of 600 *mV* (from a bandgap), the narrow pulse generator uses a current source to charge the C_0 capacitor. The comparator outputs a high voltage and the output inverter outputs a low voltage.

Then, when the current flowing into C_0 is approximately 1 μs and the capacitor voltage reaches the threshold, the comparator will output a low voltage. The output inverter gives out a high voltage. At the same time, the low side inverter receives the comparator low voltage after a delay of approximately 10 ns. Then it changes it to high voltage in order to turn on the NMOS C_1, which discharge C_0 quickly. The simulation results for narrow and ramp waves are shown in Figure 5.45.

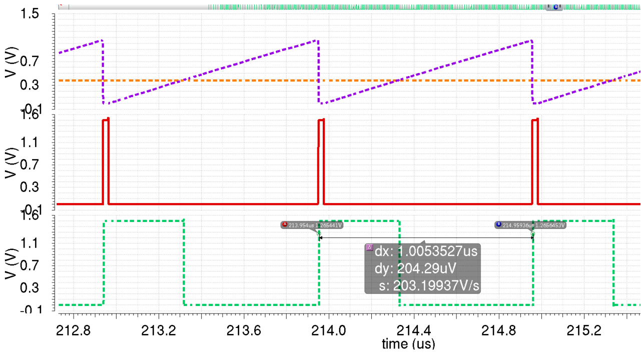


Fig. 5.45: Simulation result of the PWM wave generator

The green dotted line is the PWM signal; the red line is the narrow pulse signal; the yellow dotted line is the error amplifier output signal and the purple dotted line is the ramp signal. It can be seen that the cycle time is 1.005 μs with a deviation of 0.005 μs. This can be acceptable. In the narrow pulse simulation, the narrow pulse time is approximately 15 ns, which is slightly larger than the designed 10 ns. The difference between 5 ns should come from the inverter time. The ramp signal amplitude is 1.05 V, which is slightly less than the designed 1.1 V. The difference should come from the parasitic capacitance.

All deviations are small and can be ignored. The simulation results of the PWM generator show that the PWM part works well and meets the requirements of the buck converter.

The next simulation circuit is the dead time controller. The dead time control architecture has been discussed in the previous section. The layout of the dead time controller is in Figure 5.46.

The layout of the dead time controller occupies the area 2667 μm^2 (127 μm length multiply 21 μm width). The effect of the dead time controller is shown in Figure 5.47. Because the output PWM driver will connect to the voltage which is higher than the core voltage (1.33 V), it uses the 2.5 V I/O transistors. When 2.5 V transistors connecting the pads, it must have enough distance each other in order to avoid the latch-up problem.

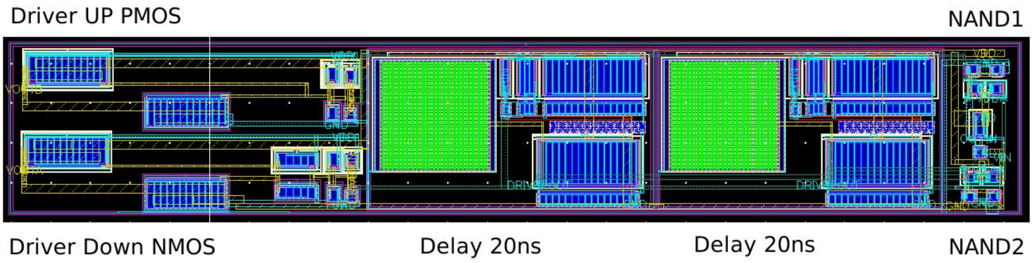


Fig. 5.46: Layout of the dead time controller

The layout of the dead time controller occupies the area 2667 μm^2 (127 μm length multiply 21 μm width). The effect of the dead time controller is shown in Figure 5.47. Because the output PWM driver will be connected to a voltage higher than the core voltage (1.33 V), it uses 2.5 V I/O transistors. When 2.5 V transistors connecting the pads, it must have enough distance between transistors to avoid latch-up problems.

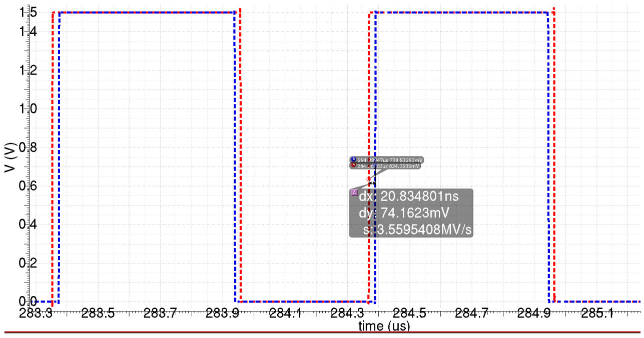


Fig. 5.47: Simulation result of the dead time controller

The red dotted line is the output voltage to control the PMOS, and the blue one is to control the NMOS. The simulation result of the dead time controller shows that the dead time is 20 *ns*, which is enough to turn on the NMOS/PMOS before switching off the PMOS/NOMS.

5.4.2 Simulation of the Error Amplifier and Compensation

Here simulates the error amplifier open-loop bode, which limits the speed of the buck converter. The error amplifier uses the same OPA architecture as used in the one used in the bandgap with OPA. The error amplifier layout is shown in Figure 5.48, which is 67 μm length and 16 μm width.

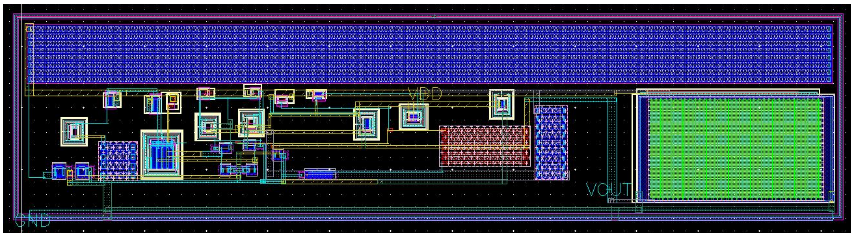


Fig. 5.48: Layout of the error amplifier

To have a better radiation aware, all key transistors use guard rings. The simulation result of the error amplifier is shown in Figure 5.49.

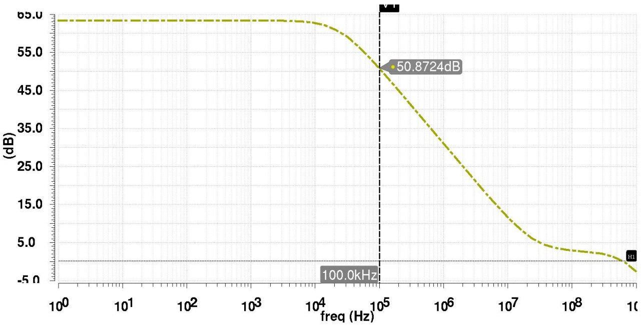


Fig. 5.49: Open-loop bode plot of the OPA

When the crossover frequency is set to 100 kHz, the open-loop gain of the error amplifier is 50 dB, which is large enough for the system closed-loop gain. When the error amplifier does not influence the closed-loop compensation, simulating the error amplifier with type III-B compensation is shown in Figure 5.50.

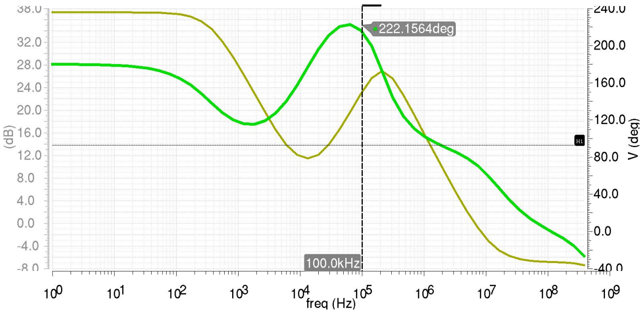


Fig. 5.50: Compensated error amplifier bode plot

The simulated Bode plot shows that due to lead compensation, the error amplifier has a large phase margin of 220° at the required crossover frequency of 100 kHz.

There is a slight deviation from the peak frequency because the lead pole F_{p2} is greater than the third pole. So F_{p3} plays a role in the lead compensation.

The whole buck closed-loop transfer function has the following Bode plot, as shown in Figure 5.51:

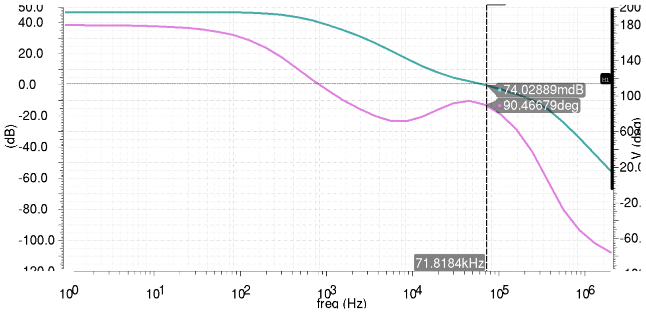


Fig. 5.51: Closed-loop bode plot of the buck converter

The Bode plot shows the crossover frequency of 78 *kHz*, which is different from the set value of 100 *kHz*. At the same time, the phase margin increases from 70 degree to 90 degree. One of the reasons is parasitic capacitance and resistance, which accelerates the amplitude drop. Therefore, the system achieves higher stability at the cost of the lower speed.

When the output current is 1 *A*, the simulation result of the output voltage is shown in Figure 5.52. The start-up time is about 100 μs.

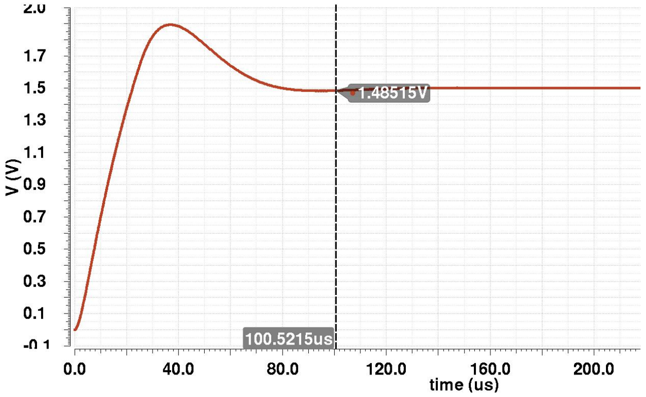


Fig. 5.52: Output voltage simulation result of buck converter with 1 *A* load

The simulation result shows that the output voltage ripple is about 2.5 *mV*. The buck converter can work normally. The whole layout of the buck converter is shown in Figure 5.53.

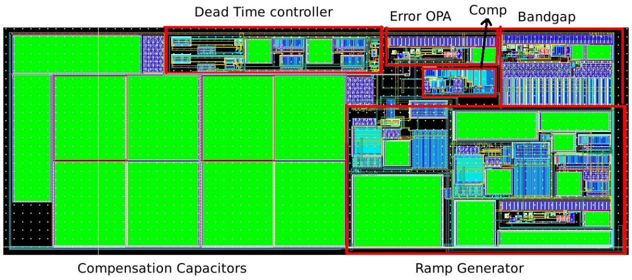


Fig. 5.53: Buck PWM generator and the compensation layout

The on-chip part is 360 μm length and 130 μm width. It requires 8 pins: the 1.5 V power supply, ground, input ramp, output ramp, voltage reference, feedback voltage, the PWM for the PMOS and the PWM for the NMOS. The on-chip part can work together with the off-chip part and output 1.5 V with an adjustable input voltage from 2.5 V to 3.3 V.

Chapter 6

Test Setup and Characterization Results

6.1 Chip for TSMC 65 *nm* Power Management IP Blocks

6.1.1 Floor Plane of the Chip

Based on the simulation of the LDO, buck DC-DC converter and bandgap reference circuit, one power management ASIC is made. The chip name is CHIPIX_LDO_BUCK and it is supported by the CHIPIX65 project. The CHIPIX65 project is developed for the LHC third generation pixel detector front-end.

The chip is implemented by TSMC 65 *nm* technology. It chooses the low power process CLN65LP with 1P9M_6x1z1u, which has one poly layer and 9 metal (Cu) layers. There is one thin metal layer (M1), six middle thickness metal layer (M2-M7), and two thick metal layers (M8 and M9). The M9 (top metal) is the thickest metal layer with the thickness of 3400 *nm*, which can bear 30.176 $mA/\mu m$. The chip size is of 0.94 $mm \times 0.94$ mm.

The chip includes three kinds of LDOs, two kinds of bandgaps and one buck converter. Three kinds of LDOs use different NMOS transistors. One use standard NMOS transistors (linear transistors) named LDO1, one uses H shape NMOS transistors named LDO2, and the other uses Enclosed Layout Transistors (ELT)

NMOS transistors named LDO3. The different NMOS transistors have different level radiation-aware. One current source is also included for the LDO.

The chip also includes two kinds of bandgaps with different structure. One does not use OPA named BG1 and the other uses OPA named BG2. Different structures will also have different radiation-aware abilities. To improve the radiation-aware, the both bandgaps use the diode-connected PMOS as the negative TC device.

The buck converter circuit is the last part in the chip. It is divided into two parts. One part on chip can offer a PWM generator and an error amplifier with III-B compensator. The other part, including power transistors, an output inductor and an output capacitors, will be placed off chip for their large volume.

The layout of the whole chip is shown in Figure 6.1.

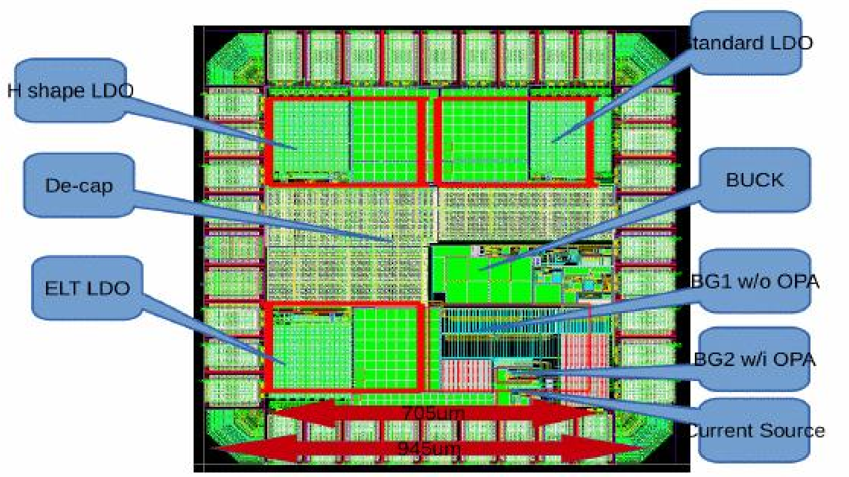


Fig. 6.1: Whole layout of the chip

Except for the area occupied by the pads, the valid area is just 705 $\mu m \times 705$ μm. And limited by the pads, the middle area cannot be used for other functions. So the space is filled by capacitors, which not only filters the input voltage, but also takes full advantage of the silicon area. There are four kinds of capacitors which can be used for filling: MOM cap, MIM cap, NMOS varactor and MOS cap. Among of them, NMOS varactor, whose density depending on the voltage, has the highest density of 5.4 $fF/\mu m^2$ at the voltage 1.5 V. So it is selected for the filtering capacitor. The structure is shown in Figure 6.2. Both N+ diffusions connect to the N-Well. And the right part of the figure shows the capacitance behavior as a function of the voltage applied at the NMOS varactor terminals.

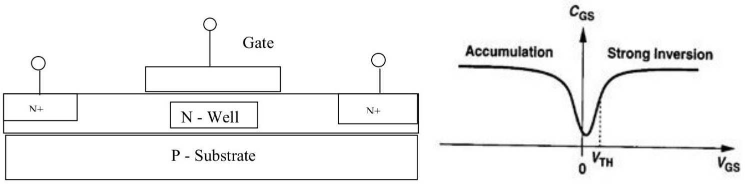


Fig. 6.2: Structure of the NMOS varactor capacitor and the C-V relationship

This capacitor is formed by a thin gate-oxide in Nwell. One terminal is the gate and the other terminal is the two ends of the N+ implant. The N+ implants have *ohm* level resistance contacting. In this design some parallel NMOS varactors can provide a capacitance of 221 *pF* capacitance for LDO1 (the standard LDO), and LDO2 (the H shape LDO), respectively. The other parallel ones can provide 215.8 *pF* to LDO3 (the ELT LDO) and the buck power supply voltage, respectively.

The technology is the CERN version of the TSMC 65 *nm* process, so the CERN pad cells are selected firstly. However, since the upper limit of the CERN pads voltage is 1.32 *V*, this is lower than the chip input voltage 1.5 *V*. Therefore, the TSMC standard I/O pad cells from the tpan65lpnv2od3 library are selected, which can bear up to 2.5 *V*. There are three kinds of 2.5 *V* I/O pads used: PVSS3A, PVDD3A, and PDB3A. PVDD3A and PVSS3A are used for the power pads VDD and GND, respectively. PDB3A is used for the signal pad. The three kinds of pads size are 50 μm width and 120 μm length. PVSS3A is shown in Figure 6.3 as an example.

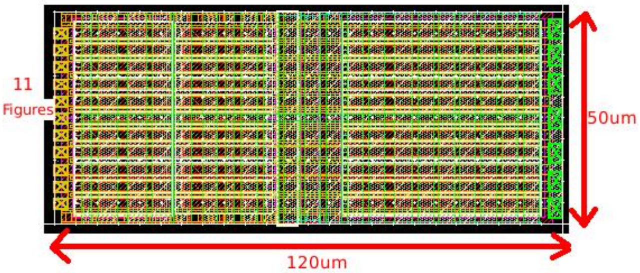


Fig. 6.3: TSMC PVSS3A pad layout

PVSS3A has 11 fingers in metal two (M2), which are used for the signal and power transferring. From the TSMC manual, the M2 current density is 1.8 $mA/\mu m$, and the pad finger width is 2.37 μm. So the maximum transferring current is:

$$I_{max,pvss3a} = 2.37\ \mu m \cdot 1.8\ mA/\mu m \cdot 11 = 46.9\ mA \tag{6.1}$$

The other two pads have different figures. PVDD3A and PDB3A have 20 and 4 fingers respectively, with the corresponding widths of 1.44 μm and 4.5 μm. Therefore, using the same calculation method, the maximum currents of PVDD3A and PDB3A are 51.8 mA and 32.4 mA, respectively. These upper limit current can transfer normal signals, but cannot transmit the large power currents such as the LDO input and output 600 mA.

The normal pads use different metal layers (M1 to M6). In this design, for the large current transmission, the chip uses the metal eight (M8) and metal nine (M9, the top layer metal). This new part will cover the original pads, as shown in Figure 6.4.

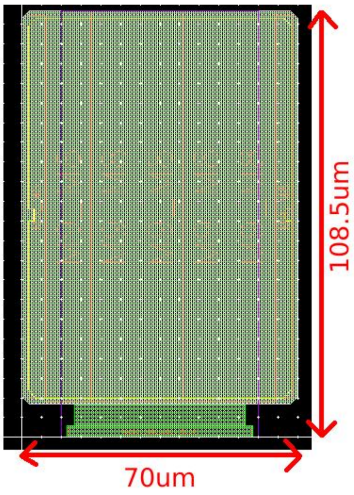


Fig. 6.4: Large current part layout covering the standard I/O pads

M8 and M9 can be connected together through some holes. Besides the electronic transmission, these holes also have a certain mechanical strength. Therefore, the holes are placed all the area between M8 and M9. When bonding the metal wires on M9, the pressure can merge the metal wires with the top metal together. When some places between the two metal layer lack the holes to support, the bonding operation may have some problems.

The M8 and M9 have the maximum current density 8 $mA/\mu m$ and 30 $mA/\mu m$, respectively. Using the same calculation method like the PVSS3A, the maximum current is:

$$I_{max,pad} = (8+30)\ mA/\mu m \cdot 70\ \mu m = 2.66\,A \tag{6.2}$$

For a maximum LDO current of 600 mA, 2.66 A is sufficient for LDO applications. A large current also generates a large amount of heat on the pads, which reduces the reliability of the wire. Therefore, large current wires will use double pads to improve reliability. In this design, some LDO input and output pads have the large currents. Therefore, the LDO input (VDD) and the output (Vout) use doubles pads. The ground pad (GND) of the internal LDO circuit requires very little current. But when this chip supplies a large current to other chips, the return current will flow back to the GND pads. Therefore, GND also requires double pads.

In this design, a new pad with M8 and M9 is required for high current circuits. For the signal path, the original pads are sufficient. The new pads have a little weaker anti-ESD protect effect than the original pads.

More pads can do more, but the distance between the two pads will be shorter. Shortened distances make bonding more difficult or impossible. Therefore, in order to easily bond the chip to the PCB, the distance is set to 18 μm. Thus every side of the chip has 9 pads, and the four sides have 36 pads together. The pads distribution is shown in Figure 6.5.

Different colors represent different devices. The light green part is LDO1; the light yellow one is LDO2; the blue one is LDO3; the green one is buck converter; the yellow one is the bandgap.

Every LDO needs 7 independent pads and together share one pad V_{ref}. Each bandgap requires one output pad, and the two bandgaps share the same power supply VDD (1.5 V) and ground (GND). The buck converter requires 8 pads for inputs and outputs. The two pads on the left are used for the anti-ESD protect ring.

The 7 pads of the LDO are dual VDD (VDD1,2,3), dual VOUT (VOUT1,2,3), dual GND (GND1,2,3) and one adjustable pad (Vadj1,2,3). All of these pads are near distribution in the power pads. The pad Vadj is used for regulate the output

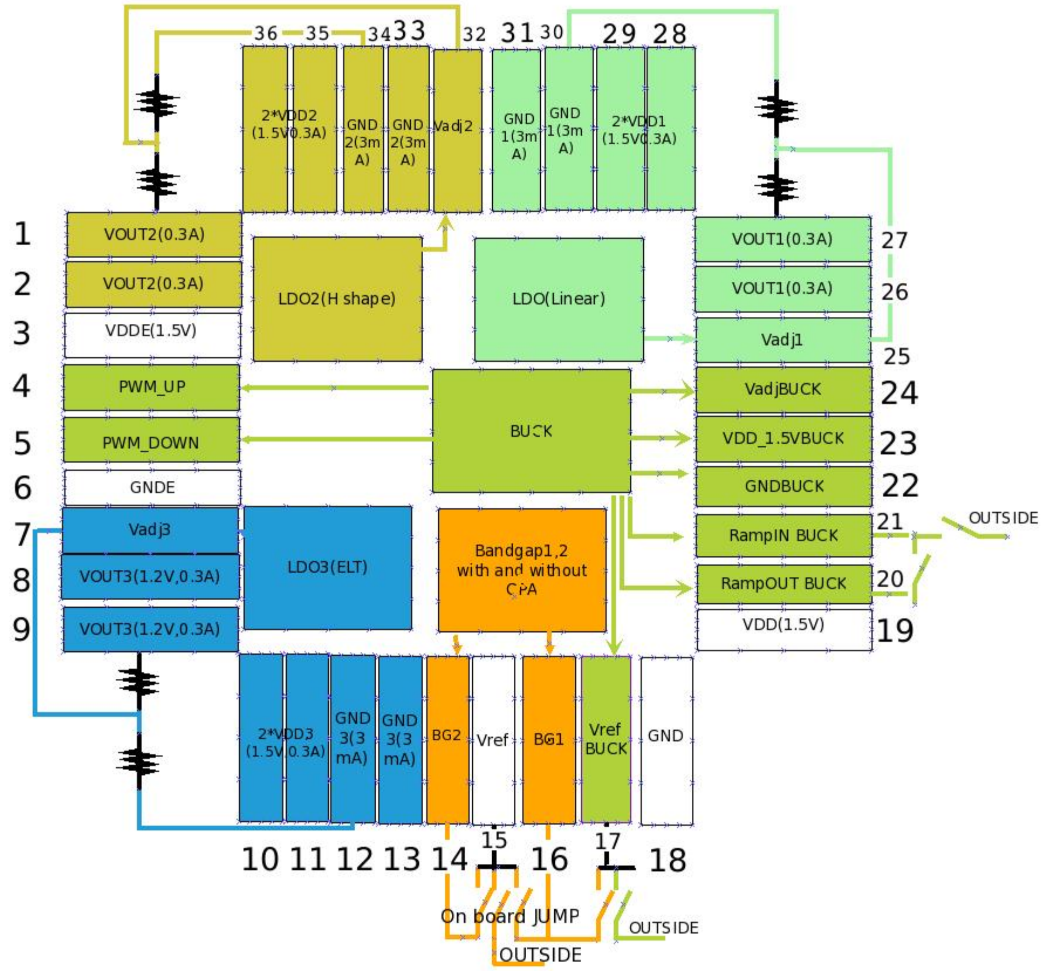


Fig. 6.5: Pads distribution of the chip

voltage, which connects the voltage divider out of the chip, which divides the VOUT and GND.

The common pad Vref of the LDO is the pad 15 at the bottom of the chip, which is close to the bandgap1 (BG1) output (pad 16) and bandagp2 (BG2) output (pad 14) to reduce the EMC interference. The off chip switch can be used to select the bandgap or the outside voltage reference to be used. The bandgap power and ground are the pad 19 and pad 18, respectively, which are close to the bandgaps to reduce the IR drop.

The buck converter has 8 pads, and the voltage reference pad Vref BUCK (pad 17) is also close to BG1 and BG2, which reduces the EMC interference and the IR drops. And there is also an off chip switch to select which bandgap or other reference voltage to use. The buck controller will output PWM signals on the PWM_UP and WMP_DONW pads, which are located on the pad 4 and 5, respectively. These two signals are relative signals. So the signal lines and the pads are put together. The ramp signals RampIN BUCK and RampOUT BUCK are located on pad 21 and pad 20, respectively. The RampIN BUCK is used to test the ramp signal situation, and the RampOUT BUCK is used to porvide the ramp signal to the comparator. The off chip switch is used to select an internal ramp signal or other ramp signal, such as a signal generator. So the two pads are put together. The buck converter internal circuit power supply VDD_1.5VBUCK (pad 23) and ground GNDBUCK (pad 22) are also placed together and close to the buck converter circuit for better EMC design. The left VadjBUCK is used to obtain feedback signal from the off chip voltage divider. The voltage divider can change the output voltage.

There are two left pads (pad 3 and 4) for the VDDE and GNDE, respectively. The both pads are used for the anti-ESD rail clamp power supply and ground, which can protect the pads from the ESD or other high voltage from outside.

In addition to the pad current limitation, another limiting path is the connection between the LDO and the power pad. The limitation comes from limited power wire widths and wire space. For example, the power connection is shown in Figure 6.6.

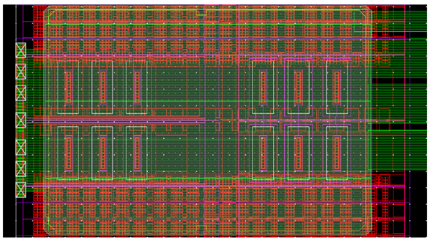


Fig. 6.6: Limited current by the power lines

Although the pad has sufficient width, the power line connecting to this pad has three 12 μm width and one 6 μm width power line, which use M9. This power line reduces the upper limit current from 2.66 A to 1.26 A. Nevertheless, it is enough for the power current, and it just has not so much current margin.

6.1.2 Anti-ESD and I/O Design Strategy

When a chip is produced or tested on board, it is often subjected to different types of electrostatic discharge (ESD) damage. Different environments produce different kinds of ESD. There are three main ESD effects: charging device model (CDM), human body model (HBM) and machine model (MM). They are shown in Figure 6.7.

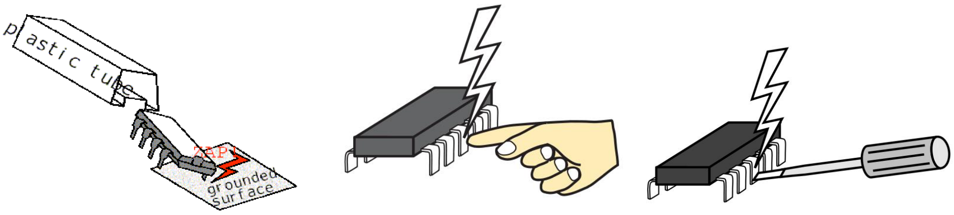


Fig. 6.7: ESD effects, the left for CDM, the middle for HBM, and the right for MM

The CDM charges come from the substrate through the internal circuit to the pads in the manufacturing environment. For example the mechanical device handling where devices slide down shipping tubes, and the test handlers that build up a charge that's subsequently discharged to ground.

The HBM charges come from the static electricity of human body. The human body can produce even several thousands volt. When human touches the chip without the anti-ESD device, the chip will be easily broken.

The MM charges come from the outside machine discharging, such as a rapid transfer of energy from a charged conductor to the conductive leads of the chip. The charged board assembly, charged cables, and the conduction arm of an automatic tester can also become the charge sources.

The current effects of three ESD models are shown in Figure 6.8.

When the technology is sub-90 *nm*, there is a new model: the ball-bonding ESD (BBD). The stress period of BBE (about 20 *ns*) is shorter than HBM and MM (about several hundred *ns*), but longer than CDM (about several *ns*). And the ESD voltage is much smaller (about 13 V) than the other three ESD models. The BBE comes from the charged wire through the pads and devices, which connect the pads and the substrate.

Many anti-ESD protection devices only can be used to protect HBM and MM, but not CDM. Because the normal ESD protection device is in the pads and there is no direct current path between the internal circuit and the anti-ESD protection device.

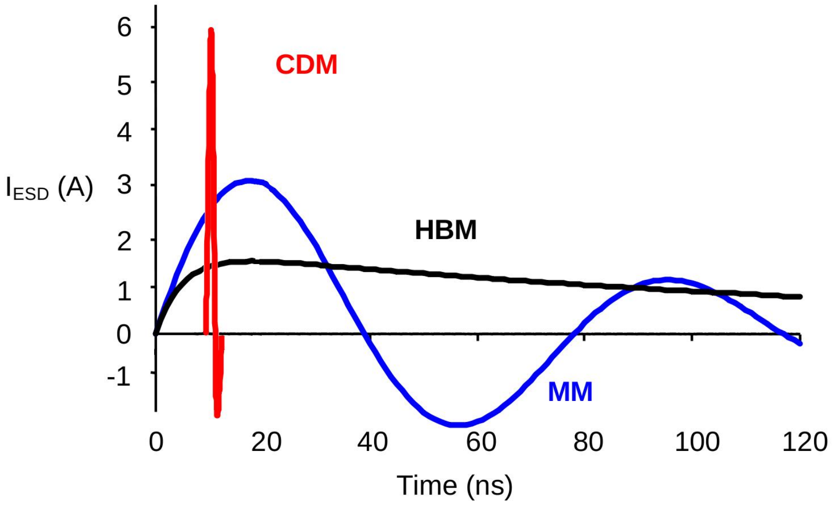


Fig. 6.8: Simulation ESD currents of CDM, HBM, and MM models under the same outside conditions

The pads are used for interacting with the externals, so all pads need the anti-ESD protection to protect the internal circuit from ESD damage, which comes from different surroundings.

There are two anti-ESD methods. One is to reduce the ESD charge and redistribute the charge through the power rail, and the other is to use the snapback effect to guide the charge to ground or the power supply. The anti-ESD network is shown in Figure 6.9.

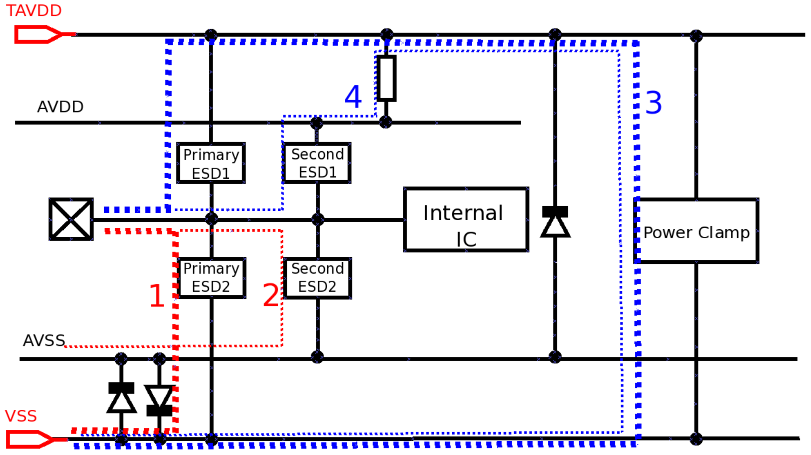


Fig. 6.9: Anti-ESD network of the chip

The first method mainly uses the diode to connect ground and the power supply, which makes any two pins have the current path through the power rail. When there is the ESD charge on the pads, some charges are induced to the power rail ring through the primary ESD1 and the second ESD2. These charges will be distributed on the power rail ring, and the extra charges will flow through the power clamp back to the VSS. Because of the limited resistor R_2, most charges flow by the path 3 (the thick blue dotted line) and a few charges use the path 4 (the thin blue dotted line).

The second method primarily uses the primary ESD2 and the second ESD2 to induce the ESD charges flowing through the path 1 (the thick red dotted line) and the path 2 (the thin red dotted line). For the same reason as the blue lines, most of the ESD charges will flue out through the primary EDS2. The primary anti-ESD and the second anti-ESD circuits can be implemented by the diode and the GGNOMS with snapback voltage.

The second ESD1 and ESD2 are useful for the ESD CDM model. In this design, the input power voltages are larger than the core voltage of 1.2V, so the power clamp between AVDD and AVSS can be ignored. However, if the core power supply is used, the power clamp is necessary. The anti-ESD network is realized in Figure 6.10.

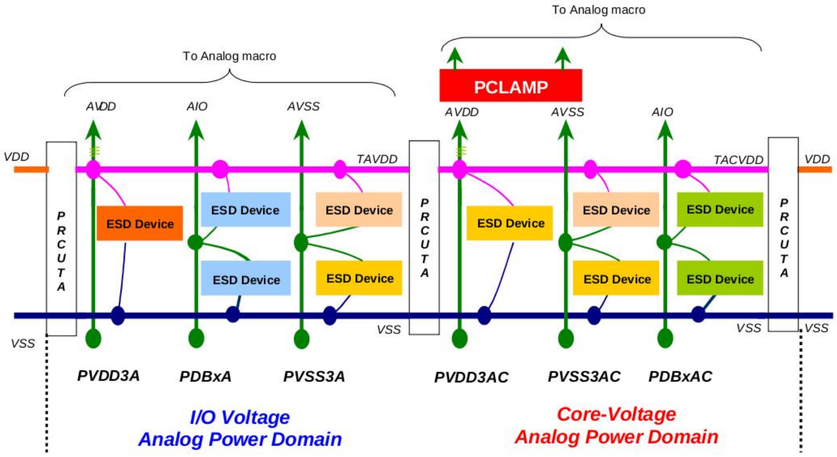


Fig. 6.10: Anti-ESD design in the pads

Figure 6.10 includes the I/O voltage and core voltage analogue power domain. The different power domains use the different pads. This design uses the left part I/O voltage power domain.

To reduce latch-up, the resistance between the pads and the power supply or ground should be less than 1 Ω. Besides the different power voltage, the different power domains have also different power rail rings, which connect all pads to the power ring. For example, the I/O and core voltage analogue power domain have only two voltage rings: VSS and TAVDD. The digital domain with the standard pads has four voltage rings: VSS, VDD, VSSPST and VDDPST.

All domains have the same VSS because they share the same substrate. Different domains are divided by the power cut analogue (PRCUTA) cell, which only allows the VSS ring going through. The analogue PRCUTA cell cuts the analogue domains, and the digital PRCUTA cell cuts the digital domains. The analogue and digital domains are cut by the analogue PRCUTA cell.

In Figure 6.9 the diodes (between AVSS and VSS) and the diodes (between TAVDD and AVSS) are realized by the PAVSS pad. The structure is shown in Figure 6.11.

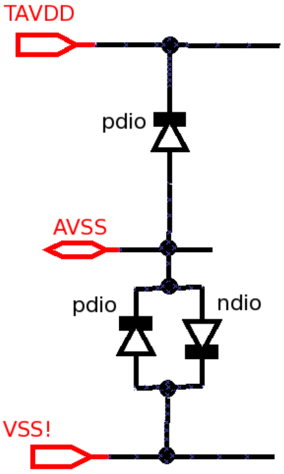


Fig. 6.11: Structure of the PVSS3A pad

When the AVSS potential is above or below VSS of a diode voltage, the ESD charge will be distributed to VSS through the bidirectional diode. AVSS also gives the ground potential of the internal circuit.

The Power Clamp in Figure 6.9 is implemented by PVDD3A. Its structure is shown in Figure 6.12. The outside gives this pad the TAVDD voltage, which is the power rail ring. Through the parallel resistors R_1 and R_2, AVDD is offered to the internal analogue macro circuits. The parallel resistance is very mall which is about 70 $\mu\Omega$, so it will not influence the AVDD voltage.

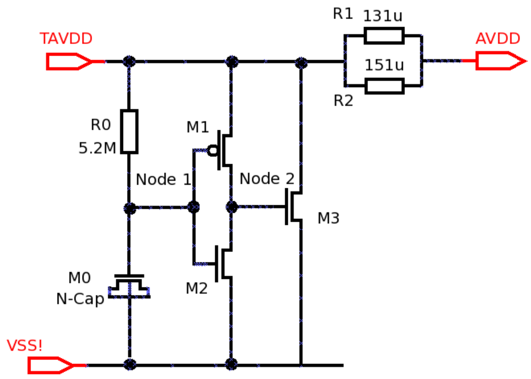


Fig. 6.12: Structure of the PVDD3A pad

M_1 and M_2 form an inverter. R_0 and C_0 form a low-pass filter with a time constant on μs level, which delays the TAVDD signal. When no ESD happens, node 1 is at high potential, so node 2 is at a low potential, which turns off the NMOS M_3. When the ESD charge is distributed to the power rail ring (TAVDD), the TAVDD potential rises. Since the low-pass filter delays the TAVDD voltage variation, the voltage crossing R_0 will be larger. There will be more current flowing through M_1 to node 2, so node 2 voltage increases, which will turn on M_3. ESD charging on TAVDD will be uncharged by M_3. So this circuit has a clamping function.

The primary ESD1 and ESD2 in Figure 6.9 are implemented by PDB3A. The structure is shown in Figure 6.13.

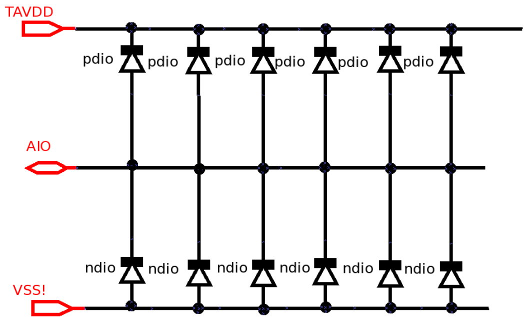


Fig. 6.13: Structure of the PDB3A pad

Six parallel diodes can not only reduce the path resistance, but also enhance the anti-static capability. It can bear 2 *kV* HBM ESD effect and has optimum design for both ESD robustness and EM (current) capacity. Some diodes are connected in series to reduce capacitance.

The second ESD1 and ESD2 are used for the CDM effect on the signal pads, implemented by internal circuitry instead of pads. It is shown in Figure 6.14.

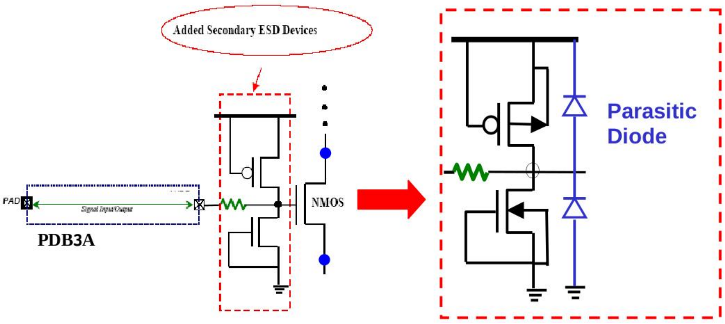


Fig. 6.14: The second anti-ESD device

When the signal input and output pads are connected to the gate terminal, a series resistor is required to protect the device. For PMOS, the resistor will connect to power, and for NMOS the resistor will connect to ground. When the signal input and output pads are connected to the drain terminal, it should have a second anti-ESD protection. The circuit uses a parasitic lateral BJT to implement a sanpback clamp through GGNMOS and GGPMOS. GGNMOS is shown in Figure 6.15 as an example, and GGPMOS has the similar structure.

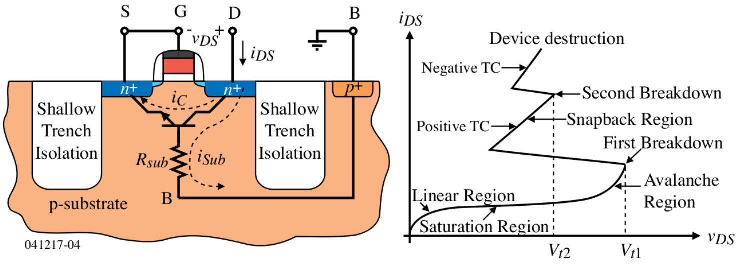


Fig. 6.15: Anti-ESD breakdown clamp of the GGNMOS

Normally, the gate of the GGNMOS is shorted to the source, so the drain current is zero. When the ESD charges accumulate on the drain, the drain voltage increases. i_{Sub} flows through R_{sub} and the triode will be turned on. At this time, the drain voltage grows up to V_{t1}. Then an avalanche happens. The triode reaches the first electric breakdown, and a lot of ESD charges will go to ground through GGNMOS. There is an important point that the current cannot go to the second breakdown point (V_{t2}). After the second point, GGNMOS will happen the heat breakdown, and cannot recover. So GGNMOS should use the multiple figures structure, and the layout ensures that the current is distributed evenly, in order to avoid the currents focusing on the few fingers, which will break GGNMOS. GGPMOS transistors have the same anti-ESD principle with GGNMOS, but there is one different. In GGPMOS, the current is formed by the holes not the electrons. The slower hole mobility makes the reaction slow down. Therefore, the anti-ESD effect of GGPMOS is weaker than GGNMOS.

The RPO layer is used for blocking the salicide layer. The salicide layer can reduce the contact resistance. But the anti-ESD circuit needs the larger drain resistance to trigger the snapback voltage easily. The internal transistors need the RPO layer increase the drain terminal resistance in order to improve the anti-ESD effect. Another method, which increases the drain terminal resistance, is to insert nwell resistors into the drain terminal.

Based on the above principle, this design avoids to use RPO resistors and nwell resistors. Because these resistors with RPO layer can be seen as the anti-ESD device, and are easily broken by the static electricity.

6.2 Electrical Test Results

In order to test the chip, this section designs one test board. It is shown in Figure 6.16. This test board can test three kinds of LDOs, two kinds of bandgaps and one buck type DC-DC respectively, and there is the fuses on the power input for safety. The J2-J6 are the jumper switches, and they are used for choosing which part will power on. To test the variable load, the triodes BC817 are chosen for the switches. The signal generator can produce the square wave, which can drive the triode BC817. The jumper J7-J9, J14 are used for connecting the transient load to the power output. The adjustable potentiometers can be used for changing the static load. The bonding board and the test board are shown in Figure 6.17.

The first figure is the bonding board, which uses the diameter 27 μm aluminum wire connecting the pins and the PCB pads. The bonding wires length is from 1 *mm* to 2 *mm*, and there is the thermal glue under the chip, which can fix the chip to PCB. The second figure is the front view of the test board and the third figure is the back view. The chip is on the back of the board. To protect the chip, there is one black

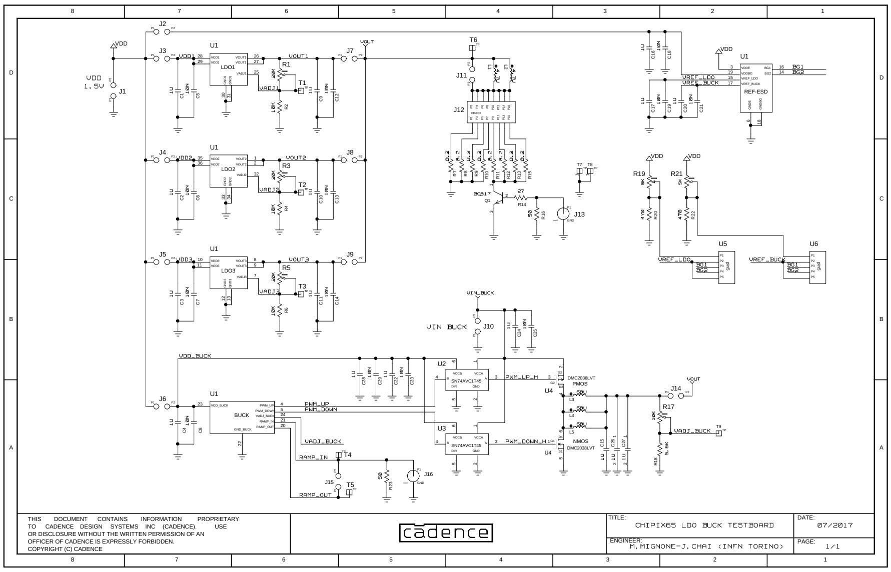


Fig. 6.16: Power management chip test schematic

cover on the chip. The forth figure is the test surrounding, which includes the power, the oscilloscope and so on.

6.2.1 LDO Electrical Tests

When testing three LDOs, BG1 is used for the reference voltage. Limited by test conditions, load regulation, line regulation, minimum voltage supply, maximum output current and transient recovery time are tested here.

Firstly, it tests the LDO load regulations. It connects the LDO output to a high-power slide rheostat. Change the output resistance, and monitor the output current. It can get the output voltage and the corresponding current. The results can be found in Figure 6.18.

The blue lines are the test lines, and the red lines are the fitting lines. When the output terminal has 30 $m\Omega$, the orange lines are the simulation lines . The LDO1 and LDO2 can get the post simulation result, but the LDO3 cannot get because of its lack the simulation model. The LDO1 NMOS is the normal transistor from the TSMC library. The LDO2 NMOS uses the H shape transistors, which can be changed on the

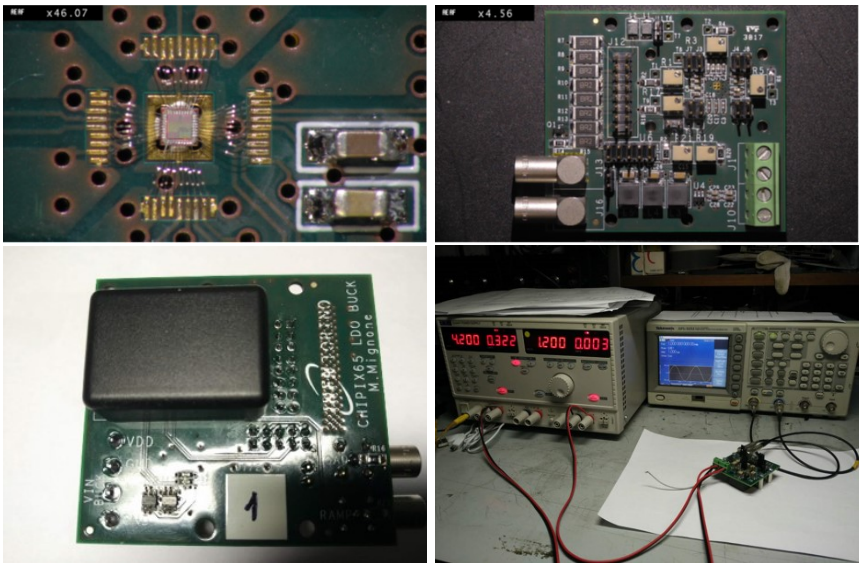


Fig. 6.17: Bonding board and test board

normal transistor, so the normal simulation model can be used. The LDO3 NMOS uses the ELT transistor, in which the S terminal surrounds the gate and the gate surrounds the D terminal. The Cadence software cannot extract parasitic parameters from ELT transistors, so ELT transistors circuit cannot do the post simulation. The normal transistor, H shape transistor and the ELT transistor increase in order the anti-radiation performance.

These three LDOs can compare different electrical properties. If there is an opportunity, we can compare their different radiation-aware performance.

The fittings in Figure 6.18 shows that the LDO1, LDO2 and LDO3 have the load regulation 40.5 mV/A, 57.5 mV/A and 64.7 mV/A, respectively. Comparing the post simulation load regulation of 1.47 mV/A, there are large differences between them. After analysis, the differences mainly come from the output line resistance. The bonding wires give out the main contribution. The bonding wire resistance can be seen in Figure 6.19. In this figure, the X-axis is the different kinds of aluminum wire diameters, and the Y-axis is how much resistance for one meter corresponding aluminum wire.

It can be seen that when using the 27 μm diameter aluminum wire, every centimeter aluminum wire has 60 $m\Omega$ and the power pad wire is about one centimeter. The power line uses the dual pads for safety, so the output lines have about 30 $m\Omega$ resistance. Based on this, when doing the post simulation, it should add 30 $m\Omega$ on

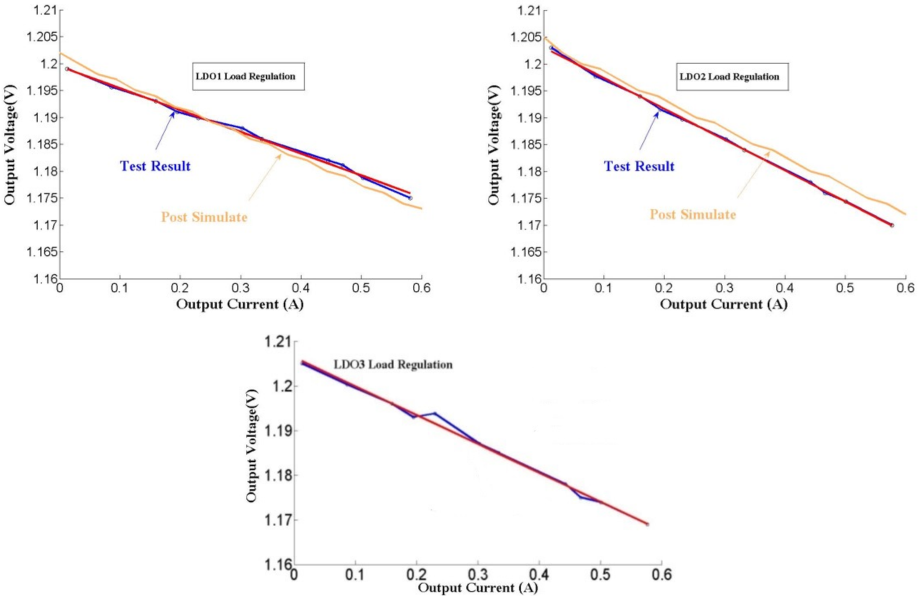


Fig. 6.18: LDO1 (top left), LDO2 (top right), LDO3 (lower) load regulations

the output. In this case, the LDO1 and LDO2's load regulations are 47.2 mV/A and 53.0 mV/A, respectively, which have the deviations 14.2% and 8.5% with the test values. These deviations are within the acceptable range.

From Figure 6.18, it can be seen that, when the output current is maximum 0.6 A, the output voltage is 1.17 V, which is slightly lower than 1.2 V. However, the output voltage can be adjusted by an external resistor, so it has little influence to the application. The test results and the post simulation with 30 $m\Omega$ can be seen in Table 6.1.

Table 6.1: LDO output load regulations table

	LDO1	LDO2	LDO3
Post-simulation result	0.0472 V/A	0.0530 V/A	
Test result	0.0405 V/A	0.0575 V/A	0.0647 V/A
Deviation	14.2%	8.5%	

LDO3 load regulation is slightly larger than LDO1 and LDO2. The reason is the LDO3 bonding wire is slightly larger, which results in a larger output resistance. However, it cannot be verified because of the lack of post simulation model.

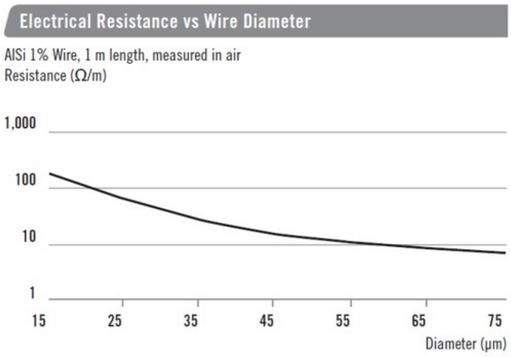


Fig. 6.19: Bonding wire resistance (Logarithmic coordinates)

Then test the LDOs's line regulations. When testing, the loads are set to 0.6 *A*. Adjust the input voltage, and monitor the output voltage variation. The results can be seen in Figure 6.20.

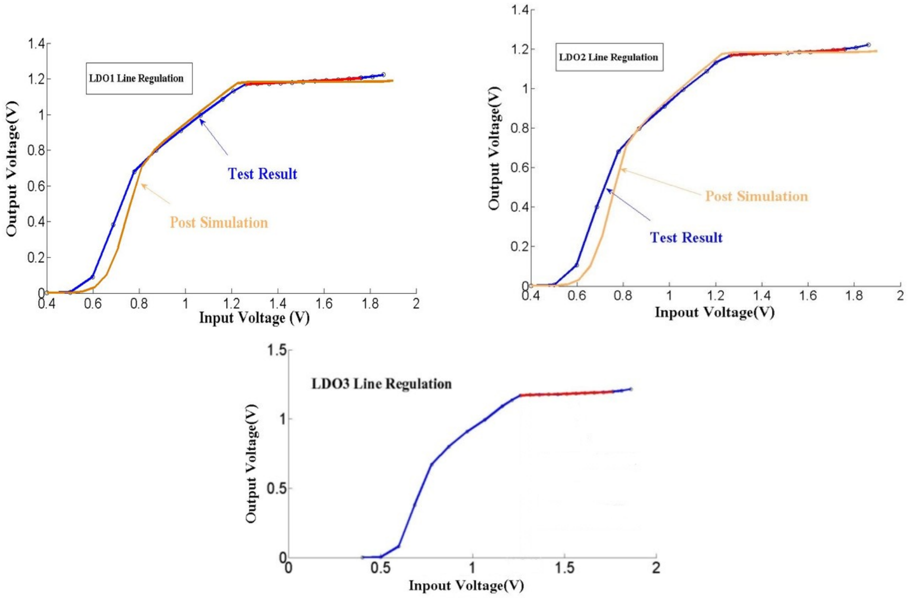


Fig. 6.20: LDO1 (top left), LDO2 (top right), LDO3 (lower) line regulations with the output current 0.6 *A*

The blue lines are the test lines, and the red lines are the fitting lines for the line regulation. The figures show that the LDO1 and LDO2 post simulation results are almost the same with the test results. For the same reason, the LDO3 cannot be compared. The three LDOs post simulation results and test results are shown in Figure 6.2.

Table 6.2: LDO output line regulations table (0.6 A)

	LDO1	LDO2	LDO3
Post-simulation result	0.0117 V/V	0.0265 V/V	
Test result	0.0691 V/V	0.0588 V/V	0.0496 V/V

Similar to the load regulation simulation results, the line regulation post simulation also has 30 $m\Omega$ on the output, which comes from the bonding wires. So the results are a little larger. In this case, the post simulation results have quite difference from the test results, as shown in Table 6.3. At the same time, the minimum power supply test results are almost the same with the post simulation results.

Table 6.3: LDO minimum power supply comparing table

	LDO1	LDO2	LDO3
Post-simulation result	1.234 V	1.294 V	
Test result	1.263 V	1.264 V	1.260 V

According to the data analysis, the large line regulation differences mainly come from the IR drop of the PCB board. To verify this conclusion, test the line regulation and do post simulation without any load.

The results can be seen in Figure 6.21. Because of the lack of LDO3 post simulation, only the LDO1 and LDO2 comparison results are shown here.

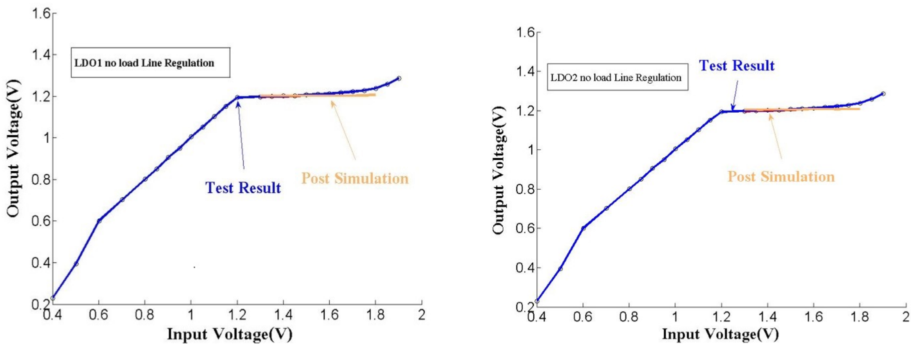


Fig. 6.21: LDO1 (left), LDO2 (right) line regulations with no load

After analysis, the no load line regulation post simulation results are shown in Table 6.4.

Table 6.4: LDO output line regulations table (0 A)

	LDO1	LDO2	LDO3
Post-simulation result	0.0041 V/V	0.0074 V/V	
Test result	0.0078 V/V	0.0134 V/V	0.0159 V/V

It can be seen that the difference between the post simulation and the test is small. This result further verifies that the PCB board and the chip IR drop can greatly affect line regulation. This is also why the three LDO test results are very different. At large current, LDO1 has a larger difference, so it can be inferred that the LDO1 input power line has a large IR drop, which reduces the LDO1 line regulation capability.

The next step is to test the transient recovery time. After the test, it can be found that the three LDOs have similar test results. Therefore, only the LDO1 oscilloscope measurement results are taken as an example here. It is shown in Figure 6.22. From the result, it can be seen that the fast load recovery time is less than 800 ns, which is 600 ns. Even though it is larger than the post simulation of 200 ns, it is a good indicator of the load transients.

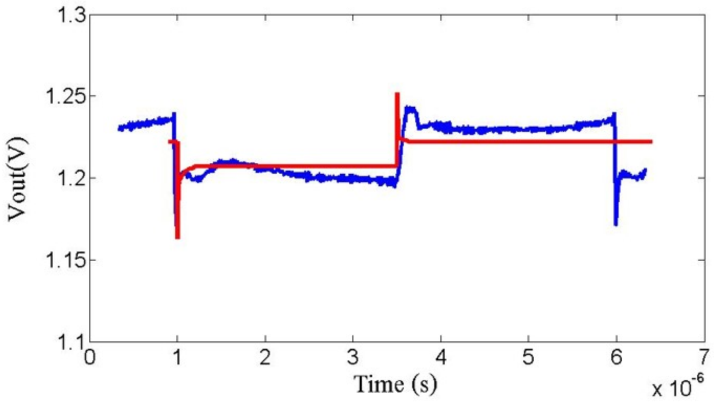


Fig. 6.22: LDO1 transient load recovery time simulation

When testing the transient recovery time, change the load current suddenly by 0.6 A, and monitor the output voltage. In Figure 6.22, the red line is the post simulation result, and the blue line is the test result. The results show that the two lines are basically the same. The test result is slightly worse than the post simulation result. The main reason is the switching action of the load current variation is slowed by the parasitic resistance and capacitance. The switching action is not fast enough, which can influence the output reaction speed. It can also be found that after the output voltage is restored, the steady voltage is a little difference with the post simulation. The difference comes from the load regulation caused by the output line resistance.

Limited by the test devices, the LDO's other parameters such as the PSR cannot be tested in current laboratory. From the current test results, some important parameter test results are not much difference from the post simulation result, and they can meet the demands of the front-end. From the horizontal comparisons between the three LDOs, it can be found that when the PMOS is the same, the three different of NMOS will have few influence on the LDO test results.

6.2.2 Bandgap Electrical Tests

The bandgap test items mainly focus on line regulation and the minimum supply voltage. Due to the lack of a temperature change chamber, it cannot perform different temperature tests.

When test the line regulation, the test measure is the same as the LDOs. The results are shown in Figure 6.23.

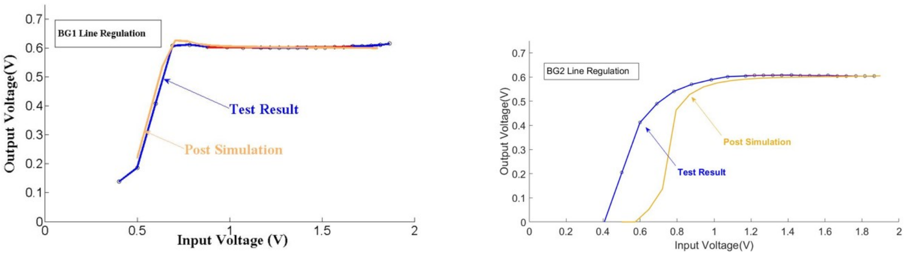


Fig. 6.23: Bandgap line regulations, left is the bandgap without OPA, right is the bandgap with OPA

BG1 is the bandgap without OPA, and BG2 is the bandgap with OPA. The blue line is the test result, the red line is the fitting line, and the orange one is the post simulation result. After the analysis, the post simulation and test results are shown in Table 6.5.

Since the bandgap input and output currents are very small, the IR drops from the parasitic resistance are also very small. From the figure, it can be seen that the post simulation and the test result of the bandgap without OPA are similar, with a deviation of less than 10%. The bandgap with OPA is better, with a deviation of

Table 6.5: Bandgap line regulations comparison table

	BG1	BG2
Post-simulation result	0.0144 V/V	0.0286 V/V
Test result	0.0160 V/V	0.0310 V/V

about 7.7%. However, when the input voltage is slowly increased, the bandgap with OPA has a larger deviation. The preliminary judgment reason is the mismatch of the MOS transistors and the resistors. This deviation dose not affect normal use, so both of the bandgaps can be used as a reference voltage. The minimum input voltage is shown in Figure 6.6.

Table 6.6: Bandgaps minimum input voltages table

	BG1	BG2
Post-simulation result	0.879 V	1.10 V
Test result	0.856 V	1.05 V

In BG1, the minimum input voltage test result is almost the same as the simulation result. In BG2, the simulation results are slightly larger than the test result, and the reason is the same as the line regulation difference. The table also shows that the bandgap without OPA is more suitable for low power supply, which can work when the input voltage less than 0.9 V. Therefore, this bandgap has an advantage in advanced technologies. The bandgap with OPA requires an input voltage larger than 1 V, which is limited by the OPA structure. So it can be used in the about 1.2 V technology. Specific to this design, both bandgaps can meet the LDO requirements.

The BG1 and BG2 start-up processes are shown in Figure 6.24.

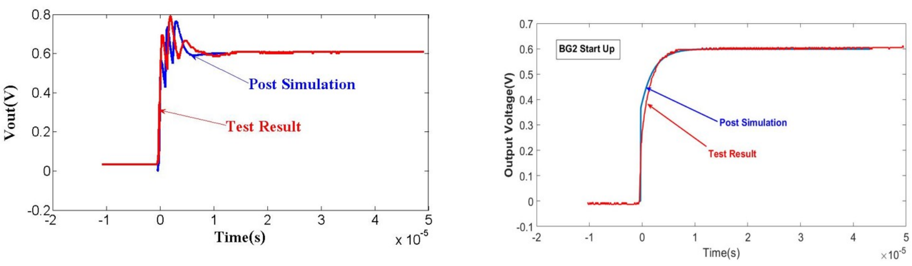


Fig. 6.24: Bandgap start-up simulation and test results, left: the bandgap without OPA, right: the bandgap with OPA

The red line is the test result and the blue one is the post simulation. The figure shows that the post simulation and the test results are almost identical. The post simulation result is of 7-8 μs larger than the simulation result in Chapter 4. The difference is, in this chapter post, simulation considering the PCB board output filtering capacitors. The total capacitance is approximately 25 pF, which includes a 20 pF filter capacitance and a approximately 5 pF parasitic capacitance between the output line and ground.

Table 6.7: Bandgap start-up time table

	BG1	BG2
Post-simulation result	10 μs	7 μs
Test result	12 μs	8 μs

BG1 and BG2 have different start-up circuits. The BG1 start-up process has some jitter. The reason is the start-up process has two switching actions (ON and OFF), which can interfere with the output voltage. And BG2 has only one switch on action, so the start-up process is more stable. But BG1 will cost less extra power consumption, because the BG1 start-up circuit will switch off after the start-up. Corresponding to this, the BG2 will always work, even if the circuit has started, so it will consume more power. Both circuits have their pros and cons. For better start-up process, it should use the BG2 start-up circuit, which will consume more power. For less power consumption, it should use the BG1 start-up circuit at the cost of longer start-up time with some jitter. For this design, there is not these limit, so both start-up circuits can meet the requirements

6.2.3 BUCK Electrical Tests

For the buck type DC-DC, the test mainly focuses on the maximum output current, the load regulation, the line regulation and the power efficiency.

The DC-DC load regulation test method is the same as the LDO and bandgap. Change the output load and monitor the output voltage. Power efficiency measurements are dependent on the ratio of output power to input power. The output power can be obtained by multiplying the output voltage and the output current. The input power comes from the power device. When the input voltage is 3.3 V, change the output load to get the corresponding power efficiency, as shown in Table 6.8.

The data in the table can be plotted in Figure 6.25. In this figure, the abscissa represents the output current and the ordinate represents the power efficiency. When the output current is 0.8 A, the power efficiency is up to 82.6%. When the load is lighter, the power efficiency decreases. The power loss mainly comes from the switching action. When the load is heavier, the power efficiency also decreases. In this case, power loss mainly comes from the MOS transistors and inductance resistance, while larger current will result in lower power efficiency. This figure also shows that the output current can be larger than 1.5 A, which is larger than the request of 1 A. To protect the chip, the test does not explore the maximum current limit.

Table 6.8: DC-DC output load variation test table

RL(Ω)	1	1.25	1.5	1.88	2.3	2.5	3	3.75	5	7.5
I_{in} (A)	0.89	0.68	0.56	0.45	0.37	0.34	0.29	0.24	0.19	0.14
V_{out} (V)	1.495	1.498	1.501	1.501	1.502	1.502	1.502	1.502	1.502	1.502
I_{out} (A)	1.49	1.19	1	0.80	0.65	0.60	0.50	0.40	0.30	0.20
Effect	75.8%	79.4%	81.2%	82.6%	81.2%	80.3%	78.4%	75.8%	71.9%	65.0%

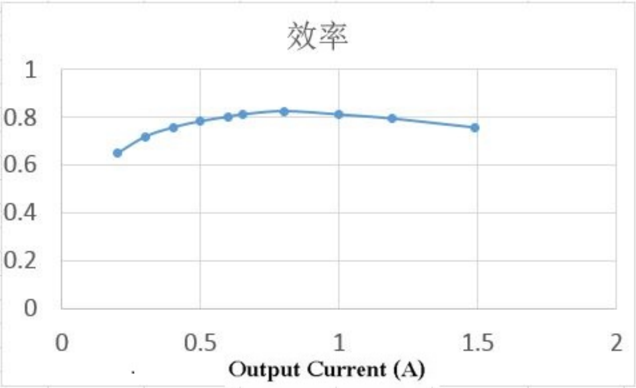


Fig. 6.25: DC-DC power efficiency figure

From Table 6.8, it can also get the relationship between output current and output voltage, which is load regulation. The result is shown in Figure 6.26. The load regulation is about 5%.

After analysis, 5% is mainly from the output line resistance, similar to the LDO load regulation.

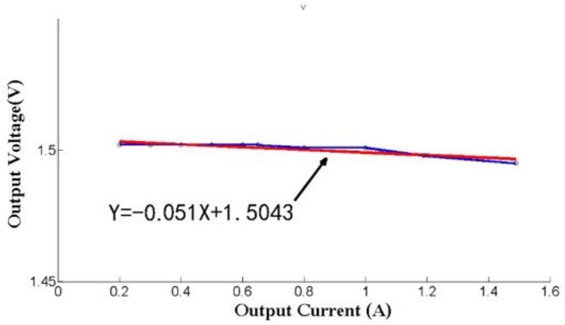


Fig. 6.26: DC-DC power load regulation

The DC-DC input voltage range is from 2.5 V to 3.3 V. The line regulation can be obtained by the similar method of LDO and bandgap. The line regulation is shown in Figure 6.27.

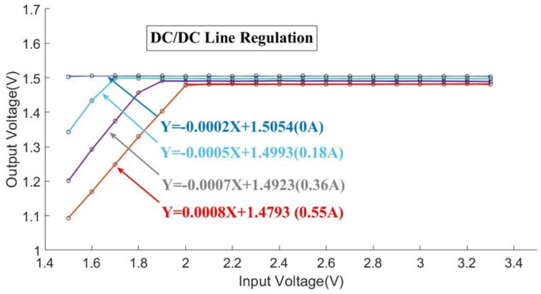


Fig. 6.27: DC-DC power line regulation

Because of the input IR drop problem, different input currents will result in different line regulations. Therefore, this test uses a different input currents: 0 A, 0.18 A, 0.36 A, 0.55 A. As can be seen from the figure, the DC-DC line regulations are 0.02%, 0.05%, 0.07%, 0.08%, respectively. The minimum supply voltages is 1.5 V, 1.7 V, 1.9 V, 2.0 V. The variation of the minimum supply voltage also comes from the IR drops. The DC-DC output voltage ripple is about 20 mV, below the request of 30 mV.

The DC-DC start-up simulation and test results are shown in Figure 6.28.

In this figure, when the load is about 0.4 A, the simulation and test results are similar. The simulation start-up time is about 100 μs, and the test start-up time is about 150 μs. The difference comes from the different switching action. The PCB parasitic capacitance and resistance can delay the test start-up.

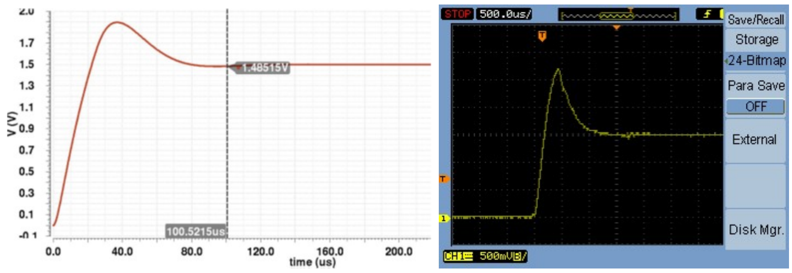


Fig. 6.28: DC-DC power start-up simulation result (left) and test result (right)

Due to the limited equipment in the laboratory, some tests cannot be achieved, including high frequency suppress situation. From the above test result, the designed DC-DC converter can meet the power requirements. When one FEE power is about 600 *mW*, the switch DC-DC converter can provide two FEE chips to operate. Therefore, the buck switch DC-DC can supply TIGER (the CGEM-IT FEE).

After the above tests, three kinds of LDOs, two kinds of bandgaps, and one buck type DC-DC can work together well under normal conditions. These blocks can form one good power management, which can provide excellent power for front-end circuits in high energy physics applications.

6.3 The Further Radiation Test

Based on the electrical test, the next step is the raditaion test. The main raidation effects are TID and SEL, so the radiation test will focus on these two effects.

To the three kinds of LDOs, the radiation test can get the different NMOS anti-radiation capability, especilly the H shape transistor. To the BESIII experiment, the 100 *krad*(Si) TID is enough, but one of the test aims is to get the maximum anti-radiation ability, so the test will be doing till the three LDOs are broken. Through the radiation-aware design, the expect TID indicator is 500 *krad*(Si).

To the both bandagps, the radiation test can evaluate the new bandgap structure anti-radiation capability, and compare the different apparance of the two bandgaps in radiation capability. The expect TID indicator is same as the LDOs.

To the buck DC-DC part, since only some circuits are on-chip, when doing the raditaion test, it is necessary to protect the off-chip circuit including the power transitors, the filter capacitor and the inductor. The test result can only reflect partly anti-radiation capability. The expect TID indicator is the same as the other blocks.

By expanding the space between transistors, and placing the substrate connectors and nwell connectors, the LDO, bandgap and DC-DC have some anti-latch up capability. The SEL test target for the blocks is 85 $MeV\text{-}cm^2/mg$, which can bear most radiation surroundings.

The radiation test will be planned to do in Padua where TID radiation testing can be performed.

Chapter 7

Conclusions and Outlook

7.1 Conclusions

This thesis covers the design and implementation of low-quiescent radiation-hard IP cores and front-end amplifier design in deep sub-micron CMOS technologies for high-energy physics applications.

A versatile power management chain based on CMOS IP cores is proposed, and includes one switching DC-DC converter, three kinds of LDOs and two kinds of bandgaps.

Two alternative low-quiescent bandgap schemes were developed and fabricated. Both circuits produce a stable 600 *mV* output with a temperature coefficient of less than 10 *ppm* between -30 °*C* to 130 °*C*.

The three LDOs use three kinds of layout structure: linear, H-shape and ELT MOSFET. Extensive previous work demonstrated the ELT as the best candidate for radiation-hardness, while an H-shape layout could be advantageous in terms of area and matching. The three flavours of LDO provide a 1.2 *V* output, and a maximum load current of 0.6 *A*. The transient recovery time is less than 1 μs.

A Buck DC-DC converter scales down the input voltage (2.5 *V*-3.3 *V*) to 1.5 *V* with a voltage ripple of about 20 *mV*, a power efficiency larger than 70%, and a maximum output current of 1 *A*.

The blocks are deployed and characterised as self-consistent CMOS IP cores, but the test-chip can be used as complete power management chain for a compact front-end board design, handling an input voltage in the range (2.5 V-3.3 V) and generating a 1.2 V voltage with a maximum output current up to 0.6 A, featuring an overall power efficiency of 60%.

A test chip including the developed IP cores was fabricated in a Muiti-project Wafer run using a TSMC 65 *nm* technology, and the electrical test results of the prototype are herein reported.

Experimental characterisation data for all circuits is coherent with post-layout simulations including the pad-ring and stray inductance of the bond wires.

7.2 Outlook

The proposed power management IP cores provide a suitable set of specifications for the use of this circuitry in state-of-the-art front-end electronics for HEP experiments. While the design of the IP cores, in terms of voltage range and output current capability, used specifications based on the future CGEM-IT on-detector electronics, the prototype is suitable to be used for power and biasing of front-end electronics modules using a power supply in the order of 1.0 V-1.2 V.

The radiation hardness of the developed IP cores is still to be assessed. Despite the fact that the lower radiation dose expected a the specific application to the CGEM-IT (10 *krad*(Si) / year), the comparison of the proposed transistor layouts under higher TID would be a strong future asset for this work. A second shortcoming of this thesis is the low power efficiency of the proposed Buck DC-DC converter.

Future research in this field should evaluate the possibility to implement a distributed power management and bias scheme on-chip. This solution, suitable to be used with linear or matrix channel/pixel chips, would use an array of dedicated bandgap and capless linear regulator on-chip. Based on the density, fill-factor and target PSRR, the regional capless LDOs and bias circuits could be used on a channel/pixel level or as sectors. Besides allowing for faster transient times at the output of the power regulator, this implementation would minimise on-chip hotspots or temperature gradients, and compensate for local power variations due to IR-drop on the supply metallisation.

References

[1] Angelo Rivetti. *CMOS: Front-end Electronics for Radiation Sensors*, volume 42. CRC Press, 2015.

[2] Willy MC Sansen. *Analog design essentials*, volume 859. Springer Science & Business Media, 2007.

[3] G. De Geronimo, D. Christian, C. Bebek, M. Garcia-Sciveres, H. Von der Lippe, G. Haller, A. A. Grillo, and M. Newcomer. Integrated Circuit Design in US High-Energy Physics. In *Proceedings, 2013 Community Summer Study on the Future of U.S. Particle Physics: Snowmass on the Mississippi (CSS2013): Minneapolis, MN, USA, July 29-August 6, 2013*, 2013.

[4] M Menouni, M Barbero, F Bompard, S Bonacini, D Fougeron, R Gaglione, A Rozanov, P Valerio, and A Wang. 1 *grad* total dose evaluation of 65 *nm* cmos technology for the hl-lhc upgrades. *Journal of Instrumentation*, 10(5):C05009–C05009, 2014.

[5] Herman Casier, Michiel Steyaert, and Arthur HM Van Roermund. *Analog circuit design: robust design, sigma delta converters, RFID*. Springer Science & Business Media, 2011.

[6] Raoul Velazco, Pascal Fouillat, and Ricardo Reis. *Radiation effects on embedded systems*. Springer Science & Business Media, 2007.

[7] Federico Faccio. Design hardening methodologies for asics. *Radiation Effects on Embedded Systems*, pages 143–160, 2007.

[8] RC Baumann. Single event effects in advanced cmos technology. In *Proc. IEEE Nuclear and Space Radiation Effects Conf. Short Course Text*, pages 1–59, 2005.

[9] Robert Baumann, Tim Hossain, Shinya Murata, and Hideki Kitagawa. Boron compounds as a dominant source of alpha particles in semiconductor devices. In *Reliability Physics Symposium, 1995. 33rd Annual Proceedings., IEEE International*, pages 297–302. IEEE, 1995.

[10] RD Schrimpf. Radiation effects in microelectronics. *Radiation Effects on Embedded Systems*, pages 11–29, 2007.

[11] Nelson S Saks, Mario G Ancona, and John A Modolo. Generation of interface states by ionizing radiation in very thin mos oxides. *IEEE Transactions on Nuclear Science*, 33(6):1185–1190, 1986.

[12] Federico Faccio. Radiation effects and hardening by design in cmos technologies. In *Analog Circuit Design*, pages 69–87. Springer, 2011.

[13] Ke-Horng Chen. *Power Management Techniques for Integrated Circuit Design*. John Wiley & Sons, 2016.

[14] David K Su, Marc J Loinaz, Shoichi Masui, and Bruce A Wooley. Experimental results and modeling techniques for substrate noise in mixed-signal integrated circuits. *IEICE transactions on electronics*, 76(5):760–770, 1993.

[15] Tallis Blalack and Bruce A Wooley. The effects of switching noise on an oversampling a/d converter. In *Solid-State Circuits Conference, 1995. Digest of Technical Papers. 41st ISSCC, 1995 IEEE International*, pages 200–201. IEEE, 1995.

[16] Xavier Aragones and Antonio Rubio. Experimental comparison of substrate noise coupling using different wafer types. *IEEE Journal of Solid-State Circuits*, 34(10):1405–1409, 1999.

[17] Mark Shane Peng and Hae-Seung Lee. Study of substrate noise and techniques for minimization. *IEEE journal of solid-state circuits*, 39(11):2080–2086, 2004.

[18] Mustafa Badaroglu, Piet Wambacq, Geert Van der Plas, Stéphane Donnay, Georges GE Gielen, and HJ De Man. Evolution of substrate noise generation mechanisms with cmos technology scaling. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 53(2):296–305, 2006.

[19] Hai Lan, Tze Wee Chen, Chi On Chui, Parastoo Nikaeen, Jae Wook Kim, and Robert W Dutton. Synthesized compact models and experimental verifications for substrate noise coupling in mixed-signal ics. *IEEE journal of solid-state circuits*, 41(8):1817–1829, 2006.

[20] Mustafa Badaroglu, Geert Van der Plas, Piet Wambacq, Lakshmanan Balasubramanian, Kris Tiri, Ingrid Verbauwhede, Stéphane Donnay, Georges GE Gielen, and HJ De Man. Digital circuit capacitance and switching analysis for ground bounce in ics with a high-ohmic substrate. *IEEE journal of solid-state circuits*, 39(7):1119–1130, 2004.

[21] Gabriel Rincon-Mora. *Analog IC design with low-dropout regulators (LDOs)*. McGraw-Hill, Inc., 2009.

[22] Vadim Ivanov. Design methodology and circuit techniques for any-load stable ldos with instant load regulation and low noise. *Analog Circuit Design*, pages 339–358, 2009.

[23] Paul R Gray, Paul Hurst, Robert G Meyer, and Stephen Lewis. *Analysis and design of analog integrated circuits, 5th Edition*. Wiley, 2009.

[24] Richard J Reay and Gregory TA Kovacs. An unconditionally stable two-stage cmos amplifier. *IEEE journal of solid-state circuits*, 30(5):591–594, 1995.

[25] Jingjing Hu, Johan H Huijsing, and Kofi AA Makinwa. A three-stage amplifier with quenched multipath frequency compensation for all capacitive loads. In *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, pages 225–228. IEEE, 2007.

[26] Young-il Kim and Sang-sun Lee. A capacitorless ldo regulator with fast feedback technique and low-quiescent current error amplifier. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 60(6):326–330, 2013.

[27] Ashis Maity and Amit Patra. Tradeoffs aware design procedure for an adaptively biased capacitorless low dropout regulator using nested miller compensation. *IEEE Transactions on Power Electronics*, 31(1):369–380, 2016.

[28] Sung-Wan Hong and Gyu-Hyeong Cho. High-gain wide-bandwidth capacitorless low-dropout regulator (ldo) for mobile applications utilizing frequency response of multiple feedback loops. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 63(1):46–57, 2016.

[29] Yoni Yosef-Hay, Pere Llimós Muntal, Dennis Øland Larsen, et al. Capacitor-free, low drop-out linear regulator in a 180 *nm* cmos for hearing aids. In *Nordic Circuits and Systems Conference (NORCAS), 2016 IEEE*, pages 1–5. IEEE, 2016.

[30] Behzad Razavi. *Design of Analog CMOS Integrated Circuits,second edition*. McGraw-Hill Higher Education, 2016.

[31] Hironori Banba, Hitoshi Shiga, Akira Umezawa, Takeshi Miyaba, Toru Tanzawa, Shigeru Atsumi, and Koji Sakui. A cmos bandgap reference circuit with sub-1 *v* operation. *IEEE Journal of Solid-State Circuits*, 34(5):670–674, 1999.

[32] Gaurav Panchanan. A sub-1 *v*, micropower bandgap reference. 2012.

[33] Guang Ge, Cheng Zhang, Gian Hoogzaad, and Kofi AA Makinwa. A single-trim cmos bandgap reference with a 3σ inaccuracy of 0.15% from -40°c to 125°c. *IEEE Journal of Solid-State Circuits*, 46(11):2693–2701, 2011.

[34] Yat-Hei Lam and Wing-Hung Ki. Cmos bandgap references with self-biased symmetrically matched current–voltage mirror and extension of sub-1 *v* design. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 18(6):857–865, 2010.

[35] A-J Annema. Low-power bandgap references featuring dtmosts. *IEEE Journal of Solid-State Circuits*, 34(7):949–955, 1999.

[36] Anne-Johan Annema and George Goksun. A 0.0025 mm^2 bandgap voltage reference for 1.1 v supply in standard 0.16 μm cmos. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pages 364–366. IEEE, 2012.

[37] Anne-Johan Annema, Paul Veldhorst, Gerben Doornbos, and Bram Nauta. A sub-1 v bandgap voltage reference in 32 nm finfet technology. In *Solid-State Circuits Conference-Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, pages 332–333. IEEE, 2009.

[38] Giuseppe De Vita and Giuseppe Iannaccone. A sub-1 v, 10 $ppm/^\circ$c, nanopower voltage reference generator. *IEEE Journal of Solid-State Circuits*, 42(7):1536–1542, 2007.

[39] G Giustolisi, G Palumbo, M Criscione, and F Cutri. A low-voltage low-power voltage reference based on subthreshold mosfets. *IEEE Journal of Solid-State Circuits*, 38(1):151–154, 2003.

[40] P Kinget, Christos Vezyrtzis, Ed Chiang, B Hung, and TL Li. Voltage references for ultra-low supply voltages. In *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, pages 715–720. IEEE, 2008.

[41] Haruo Kobayashi and Takashi Nabeshima. *Handbook of Power Management Circuits*. CRC Press, 2016.

[42] Naeim Safari. Design of a dc/dc buck converter for ultra-low power applications in 65 nm cmos process, 2012.

[43] Zhipeng Li et al. *Design of a step-down DC-DC controller integrated circuit with adaptive dead-time control*. PhD thesis, Massachusetts Institute of Technology, 2010.

[44] Doug Mattingly. Designing stable compensation networks for single phase voltage mode buck regulators. *Intersil Technical Brief*, pages 1–10, 2003.

[45] Robert W Erickson and Dragan Maksimovic. *Fundamentals of power electronics*. Springer Science & Business Media, 2007.

[46] Amir M Rahimi, Parviz Parto, and Peyman Asadi. Compensator design procedure for buck converter with voltage-mode error-amplifier. *International Rectifier Co Application Note, AN-1162*, 2010.

[47] R Debbe, J Fischer, D Lissauer, T Ludlam, D Makowiecki, E O'Brien, V Radeka, S Rescia, L Rogers, GC Smith, et al. A study of wire chambers with highly segmented cathode pad readout for high multiplicity charged particle detection. *IEEE Transactions on Nuclear Science*, 37(2):88–94, 1990.

[48] BESIII Collaboration. BESIII Cylindrical GEM Inner Tracker. Technical report, July 2014.

[49] Lia Lavezzi, M Alexeev, A Amoroso, R Baldini Ferroli, M Bertani, D Bettoni, F Bianchi, A Calcaterra, N Canale, M Capodiferro, et al. The cylindrical gem inner tracker of the besiii experiment: prototype test beam results. *arXiv preprint arXiv:1706.02428*, 2017.

Appendix A

A Custom IC Dsign for the Readout of the BESIII CGEM-IT Detector

A.1 The CGEM-IT Project

The CGEM-IT project is used in the BESIII experiment carried out at BEPC II. It is shown in Figure A.1.

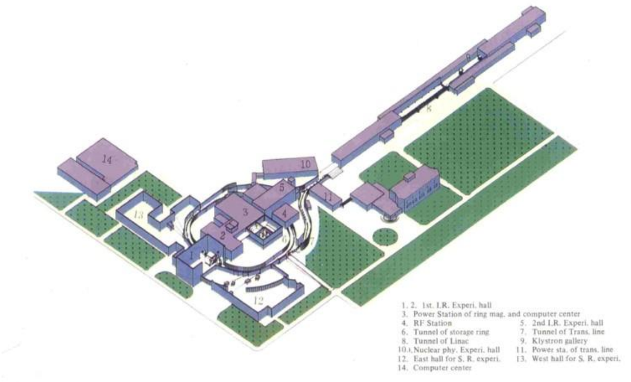
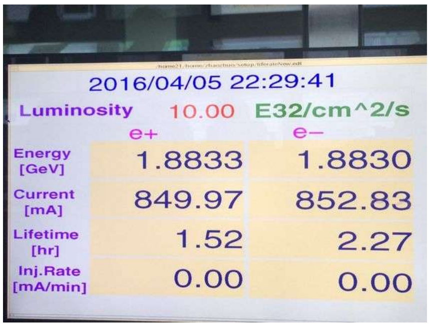

1, 2. 1st. I.R. Experi. hall
3. Power Station of ring mag. and computer center
4. RF Station
5. 2nd I.R. Experi. hall
6. Tunnel of storage ring
7. Tunnel of Trans. line
8. Tunnel of Linac
9. Klystron gallery
10. Nuclear phy. Experi. hall
11. Power sta. of trans. line
12. East hall for S. R. experi.
13. West hall for S. R. experi.
14. Computer center

Fig. A.1: Layout of BEPC

The Beijing Electron-Positron Collider II (BEPC II) is a Chinese electron-positron collider, a large particle accelerator, which is an update upgrade of the former BEPC. BEPC was built on 1989, and the update started in 2004. The target of BEPC II was a the luminosity 100 times better than BEPC. This target has been realized on April 5, 2016, as shown in Figure A.2.



2016/04/05 22:29:41

Luminosity 10.00 E32/cm^2/s

	e+	e−
Energy [GeV]	1.8833	1.8830
Current [mA]	849.97	852.83
Lifetime [hr]	1.52	2.27
Inj.Rate [mA/min]	0.00	0.00

Fig. A.2: The word record of BEPC II

The energy range of BEPC II mainly produces some high energy particles including τ, J/ψ and its related particles, so BEPC II has another name "the factory of τ, J/ψ ".

Along with the update of BEPC II, the main detector was also updated from BES to BESIII. BESIII detects the high energy particles produced by the collisions of high energy electron-positron. BESIII is shown in Figure A.3.

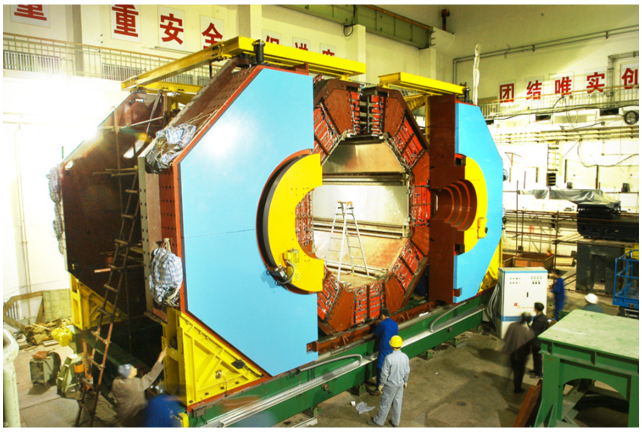


Fig. A.3: BESIII actual picture

BESIII consists of four main detectors: the Main Drift Chamber (MDC), the Time Of Flight system (TOF), the Electro-Magnetic Calorimeter (EMC) and the Muon Chamber (MUC). The different detectors are shown in Figure A.4.

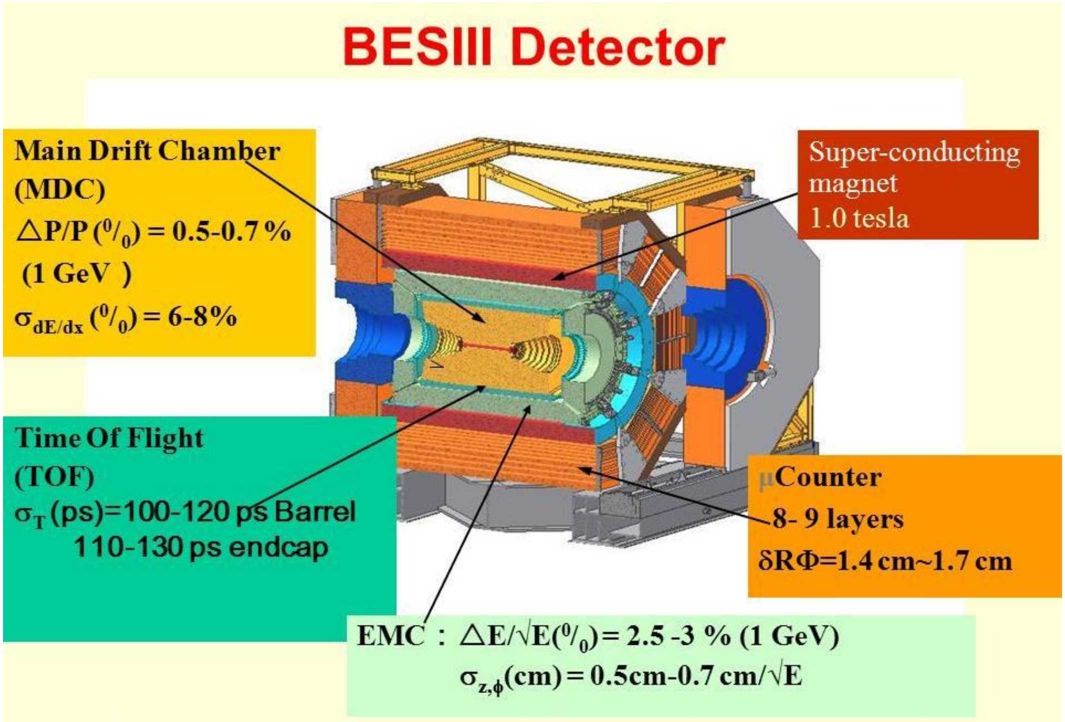
The TOF purpose is to measure the flight time of charged particles for particle identification (PID).

The EMC primary function is to precisely measure energies and positions of electrons and photons.

The MUC (muon counter) is a gaseous detector based on Resistive Plate Chambers (RPCs).

The Main Drift Chamber is the innermost sub-detector of the BESIII detector. It is one of the most important sub-detectors. Its main functions are:

- Precise momentum measurement. To achieve this, special cares should be taken to minimize the effects of multiple Coulomb scattering in the design;
- Adequate dE/dx resolution for particle identification;
- Good reconstruction efficiency for short tracks from interaction point;
- Realization of charged particle trigger at level one;
- Maximum possible solid angle coverage for charged track measurement.



BESIII Detector

Main Drift Chamber (MDC)
$\Delta P/P\,(^0/_0) = 0.5\text{-}0.7\,\%$
(1 GeV)
$\sigma_{dE/dx}\,(^0/_0) = 6\text{-}8\%$

Super-conducting magnet
1.0 tesla

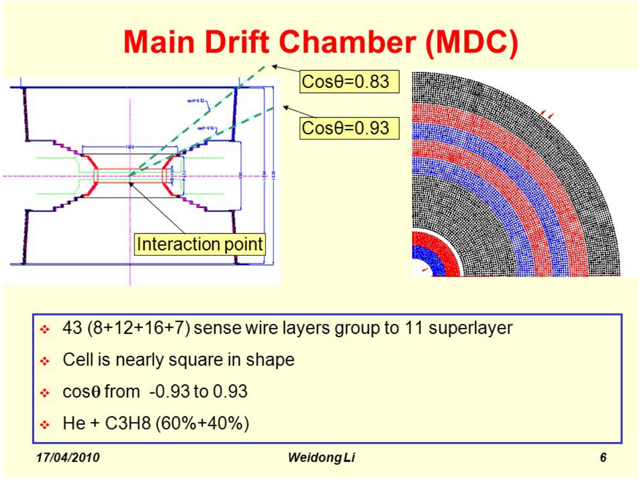
Time Of Flight (TOF)
σ_T (ps)=100-120 ps Barrel
110-130 ps endcap

μCounter
8- 9 layers
$\delta R\Phi$=1.4 cm~1.7 cm

EMC : $\Delta E/\sqrt{E}(^0/_0) = 2.5\text{ -}3\ \%$ (1 GeV)
$\sigma_{z,\phi}(\text{cm}) = 0.5\text{cm-}0.7\ \text{cm}/\sqrt{E}$

Fig. A.4: Detectors of BESIII

The MDC has 43 layers, which are divided into two trackers, and the trackers share the same He based gas mixture. The inner tracker includes 8 stereo layers. From inside to outside, the outer tracker includes the 12 axial layers, 16 stereo layers, and 7 axial layers. The MDC is shown in Figure A.5.



Main Drift Chamber (MDC)

- 43 (8+12+16+7) sense wire layers group to 11 superlayer
- Cell is nearly square in shape
- cosθ from -0.93 to 0.93
- He + C3H8 (60%+40%)

17/04/2010 Weidong Li 6

Fig. A.5: MDC architecture of BESIII

As time goes by, the MDC meets some problems: there is the significant anode aging effect in the inner tracker, and the increased luminosity will speed up the anode aging. The anode aging results in a lower efficiency, the aging rate is about 4% per year. It is shown in Figure A.6.

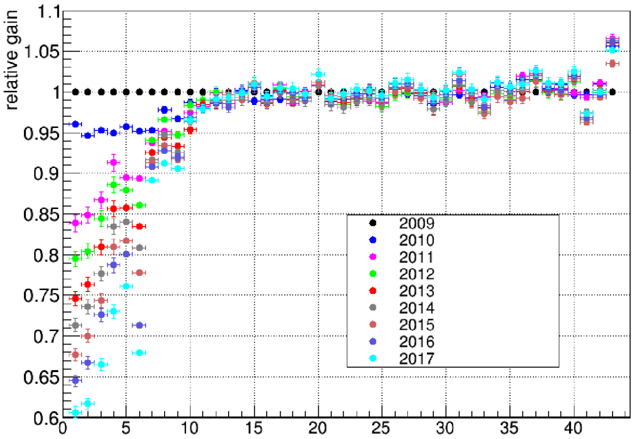


Fig. A.6: Anode aging of the MDC (up to 2017)

In this situation, the inner tracker needs to be replaced in order to run up to 2022 or more. The Italian cooperation group proposed that a CGEM detector can be used to replace the inner tracker. The CGEM (Cylindrical GEM) detector based on GEM detectors, has been already used in the KLOE-2 experiment successfully. This replace MDC project is the CGEM-IT which is the abbreviation of Cylindrical Gas Electron Multipliers Inner Tracker.

A.2 Readout System of the CGEM-IT

The CGEM-IT detector has three layers, and each layer is composed by a triple cylindrical GEM. It is shown in Figure A.7.

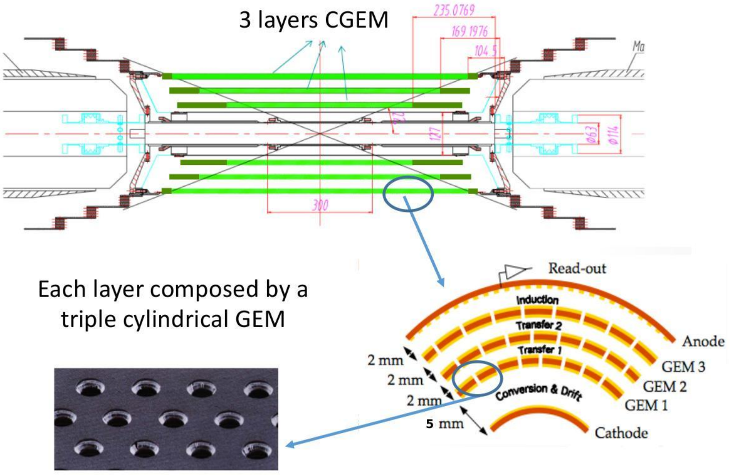


Fig. A.7: Three layers architecture of CGEM-IT

In the CGEM-IT, it is impossible to read out every position signal separately at the present technology level. To reduce the number of readout channel, this project uses the two-dimensional readout method. The anode readout picture is shown in Figure A.8.

The X strips are one dimensional, and the V strips are the other dimensional. The two kind of strips form the α angle. The hit positions can be located by the two direction strips. The goal of spatial resolution (the pitch) is 130 μm, so normally the distance between microstrips is 130 μm. But in this case, there is also too many channels to be read and it is difficult to read them. In this design, the CGEM-IT can use a larger pitch to get the better space resolution. The specific methods are described in the next section.

In the V and X directions, the pitches are 650 μm. This project uses the jagged strip for improvement, which can reduce the inter strip capacitance about 30%. The lower parasitic capacitance is useful to lower the cross talk and the ENC noise.

These strips can give a signal on the two sides of the CGEM. Due to diffusion, the charge cloud collected on the readout board is larger than the strip width. So the weighting method is used for calculating the exact track position in two dimensions.

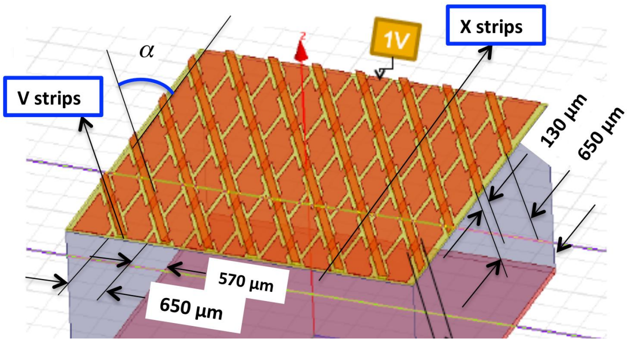


Fig. A.8: Two-dimensional readout method

The three layers respectively have different strip numbers and different angles, and the details are shown in Table A.1.

Table A.1: Different layer strips numbers

	Layer 1	Layer 2	Layer 3
X strips	846	1281	1692
V strips	1176	2193	2838
Total strips	2022	3474	3530
Stereo angle	45.9	33.1	33.0

The total channels number is 10026. These channels will transfer the signal to the two sides through the Anode flexible PCB. Take the second layer for an example, which is shown in Figure A.9.

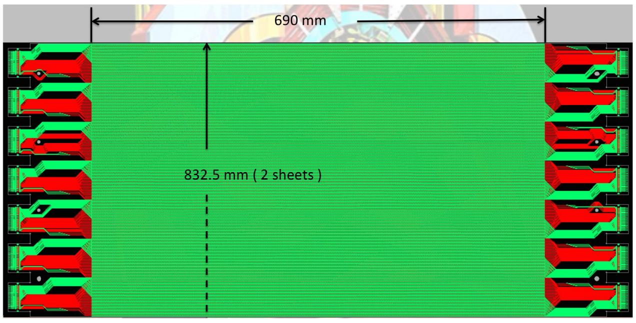


Fig. A.9: Anode PCB board of the second layer

The two PCB boards can output all signal from the second layer. The second layer has together 3474 channels, each PCB board transfers 1737 channels. If every readout terminal has the 128 channels read-out ability, each side needs 14 terminals to read out. The installed readout terminal is shown in Figure A.10.

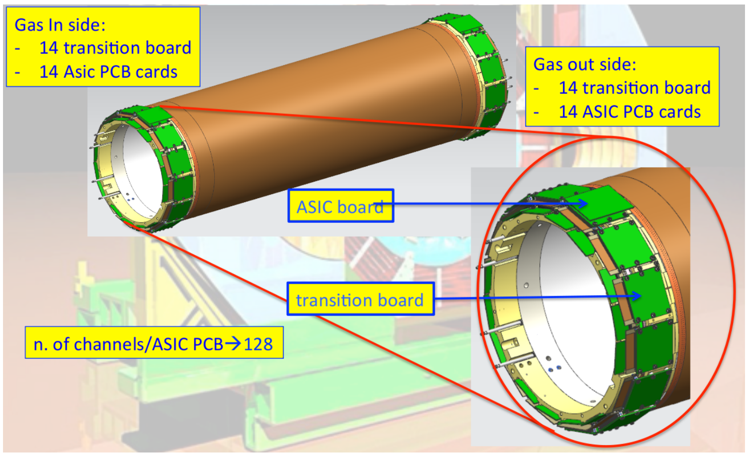


Fig. A.10: Signal readout terminal cards of the second layer

The transition board provides interconnections to the anode strips and the GND, provides support and fixing for the front-end electronics (FEE) board, and hosts the ASIC input protection network. The FEE boards offer two 64 channels ASICs, power supply and the input/output connectors.

Together all the three layers need 10026 channels, and every FEE board contains 128 channels (2 × 64channels). The whole signal readout system is shown in Figure A.11.

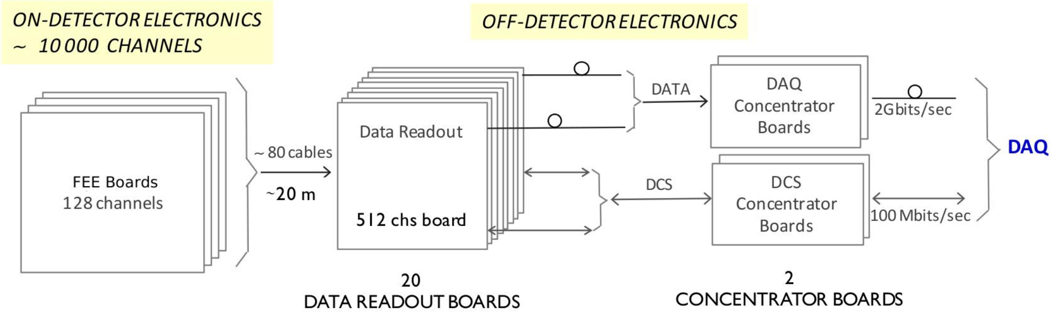


Fig. A.11: Signal readout system of the CGEM-IT

The on-detector electronics (the FEE boards) get the signal with good S/N ratio, and the off-detector electronics (the data readout boards and connection boards) dispose the data further on the experiment platform, which needs at least 20 meters long cables to connect.

A.3 Detector Conditions and FEE Indicators

The detector function decides what kinds of parameters are needed by the front-end ASIC. The main aim of the CGEM-IT is the spatial resolution $\sigma_{\mathrm{xy}} = 130\ \mu m$. To get so high spatial resolution using the 650 μm pitch is a challenge. The CGEM-IT uses the charge centroid and micro TPC together to reach the goal.

The charge centroid readout method, widely used, is based on the center-of-gravity of the induced charge distribution. The working principle of the charge centroid is shown in Figure A.12, and the position can be given using the following equation:

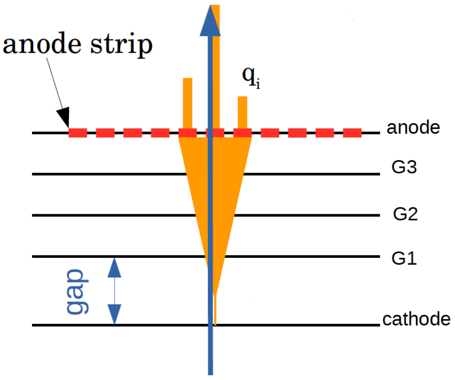


Fig. A.12: Illustration of the charge centroid for path through tracks with magnetic field B off

$$< X >= \frac{\Sigma_{\mathrm{i}} X_i q_{\mathrm{i}}}{\Sigma q_{\mathrm{i}}} \tag{A.1}$$

The position resolution of the charge centroid method has been carried out by Debbe, Radeka and O'Brien [47]:

$$\sigma = \frac{\sqrt{(n-1)} \cdot W \cdot ENC}{\Sigma_{i=1}^{n} Q_i} \tag{A.2}$$

where W is the pitch width; n means the fired strips; Q_i means charge collected by the strip i; and ENC means the Equivalent Noise Charge [48]. According to this equation, to get the 130 μm position resolution, if the input charge is 1 fC, and the fired strips number is 3, the ENC should be less than:

$$ENC \leqslant \frac{1fC \cdot 130\ \mu m}{\sqrt{(3-1)} \cdot 650\ \mu m} = 850\ e \tag{A.3}$$

This number is the whole system ENC noise, which includes all the front-end noise and the ADC noise. When the system uses the 10-bit Wilkinson ADC, the valid bits are about 9-bit which will correspond from 1 fC to 50 fC. So every step means:

$$Charge_{step} = \frac{50\ fC}{2^9} = 0.097\ fC = 610\ e \tag{A.4}$$

The ENC noise results from the ADC is:

$$ENC_{ADC} = \frac{Charge_{step}}{\sqrt{12}} = 176\ e \tag{A.5}$$

The ENC left for the analogue front-end ENC_{FEE} is:

$$ENC_{FEE} = \sqrt{ENC^2 - ENC_{ADC}^2} = 832\ e \tag{A.6}$$

This charge centroid method is suitable for the charge Gaussian distribution. When the particle crosses the GEM with a large incident angle, or there is a strong magnetic field, the charge distribution will not exactly a Gaussian. The resolution of the charge centroid will be worse. The situations are shown in Figure A.13.

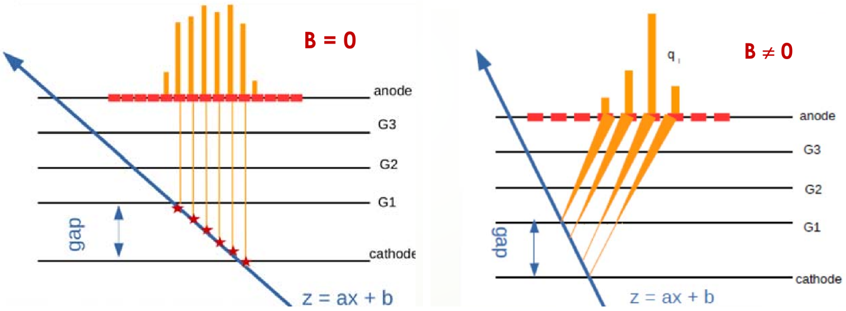


Fig. A.13: Worse situations by the charge centroid

In these situations, the micro-TPC (μTPC) method is used [49]. In this case, the drift gap can be seen as a micro time projection chamber and the position of each primary ionization z_i can be gotten by the drift velocity (about 30 $\mu m/ns$ in the MDC gas surrounding) of the electrons and the signal time on the strip.

In order to minimize the error, the middle of the drift gap (5 *mm*) corresponding x_{charge} position as a reference, which can be gotten by the charge centroid. The relative position is:

$$\Delta x = \frac{\frac{gap}{2} - b}{a} \tag{A.7}$$

So the position x can be gotten:

$$x = x_{charge} + \Delta x \tag{A.8}$$

The position resolution σ_x is:

$$\sigma_x = \sqrt{\sigma^2_{x_{charge}} + \sigma^2_{\Delta x}} \tag{A.9}$$

$\sigma_{x_{charge}}$ can be ignored, so the σ_x depends on the $\sigma_{\Delta x}$:

$$\sigma_{\Delta x} = \sqrt{\left(\frac{\sigma_a}{\overline{a}}\right)^2 + \left(\frac{\sigma_b}{\overline{b}}\right)^2} \tag{A.10}$$

The a and b can be gotten from fitting the sample points (normally from 5 points to 10 points). When the incident angle is small, $\sigma_a/\overline{a}$ and $\sigma_b/\overline{b}$ both are larger, so μTPC is more effective when the incident angle is larger.

When the charge centroid and the μTPC methods work together, there are the effects shown in Figure A.14.

From this figure, the two methods are complementary. The resolution resulting from μTPC mainly depends on the Front-end time resolution, the jitter. Through the simulation and comparing with other similar CGEM detectors, when the front-end jitter is about 5 *ns*, μTPC can have the desired resolution.

In addition, the CGEM-IT should cope with an event rate of 10^4 Hz/cm^2. Through the Monte Carlo (MC) simulation and comparing with the MDC inner tracker, the event rate of one strip is 14 *kHz*/strip. Considering 4 times safety margin, the maximum event rate on one strip is about 60 *kHz* with the tolerable loss rate 5%@60 *kHz*.

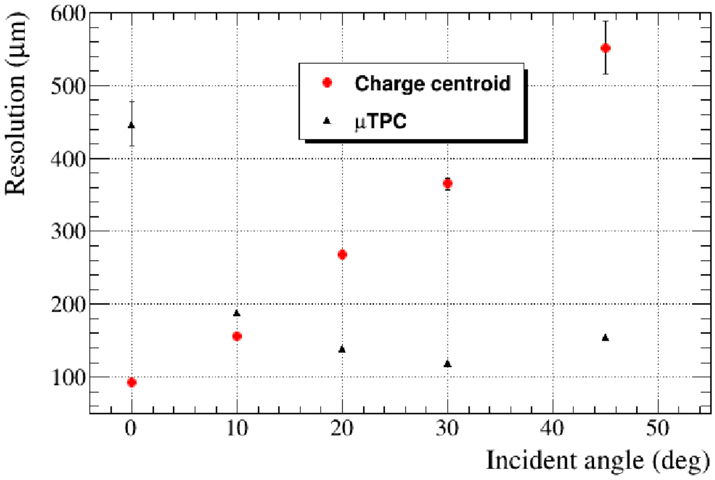


Fig. A.14: Combined resolution with different incident angles

The detector demands the data rate 60 kHz. This data rate and the tolerate loss rate can determine the dead time of the front-end design. The system dead time decides the loss rate, so the shorter the better. There are two methods to count the dead time. One is accurate method, which use the event Poisson distribution:

$$P(N,\Delta T) = \frac{(\overline{n}\Delta T)^N}{N!} e^{-\overline{n}\Delta T} \tag{A.11}$$

According to the Poisson distribution, there is the following dead time versus the event rate Figure A.15. When the tolerable pile up is 5%, the 60 kHz corresponds to the dead time of 1 μs.

The other method to get the dead time is an approximate method. When one event appear, the following dead time is T_d. And then in this dead time, there will be the $T_d \cdot 60\ Hz$ event lost. So the detection efficiency is:

$$Rate_{effi} = \frac{1}{1 + T_d \cdot 60\ Hz} \tag{A.12}$$

The numerator is one event which has been detected, the denominator means in the detecting time, how many useful events happen. When the loss rate is 5%, the effective rate is 95%, so the dead time T_d can be easy to decide. Here just considers one event caused by the dead time, and ignores two, three or more events, which are rare. So it is an approximate method.

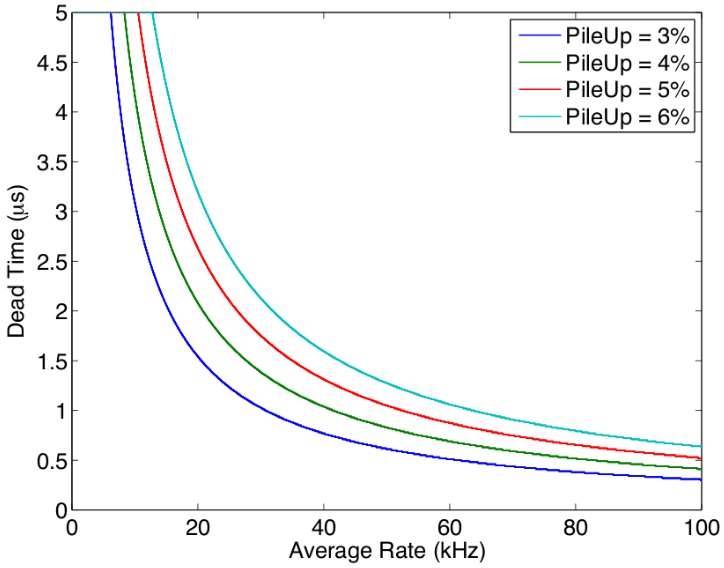


Fig. A.15: Dead time versus the rate for different pile-up probabilities

There is a parasitic capacitance between the strips and the ground, and because of the different strip length (especially the V direction), the capacitance is different accordingly. The jagged strip are used to reduce the parasitic capacitance of about 30%. Through simulation and measurement, the present max parasitic capacitance is about 50 pF. Considering the aging margin and other unknown deviation, the safety factor 2 has been used. So the expected c capacitance output the FEE is 100 pF.

The desired efficiency is about 98%, and the efficiency depends on the triple GEM gains, which results from the high voltage. The relationship between the gain and the efficiency is shown in Figure A.16. It can be seen that when the gain reaches 6000 times, the efficiency is about 97%, and the trend is flat. The higher gain needs larger voltage, which can result in the higher discharge phenomenon. So the 6000 times not only keeps the high efficiency but also has a low discharge.

Besides this benefit, the gain of the CGEM influences the resolution by the charge centroid method, as shown in Figure A.17. When the gain is about 6000, there is a better spatial resolution. The gain can influence the total charge and the fired strip.

Here there is one point which should be noticed. When the gas detection efficiency is 97% and the FEE dead time reducing detection efficiency of 5%, the whole efficiency will drop to 92%. So there is the 6% gap with the desired efficiency 98%.

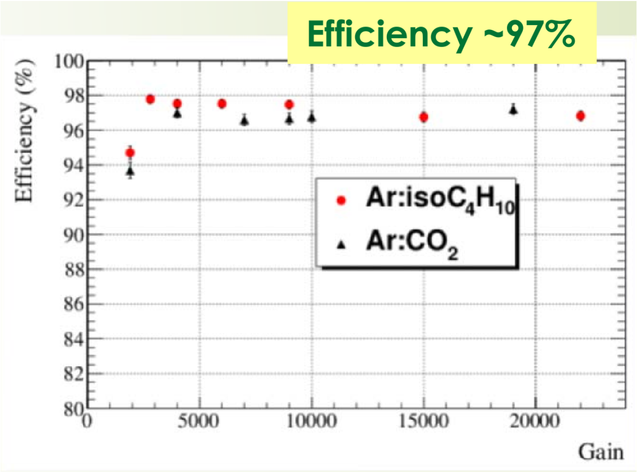


Fig. A.16: Relationship between the gain and the efficiency

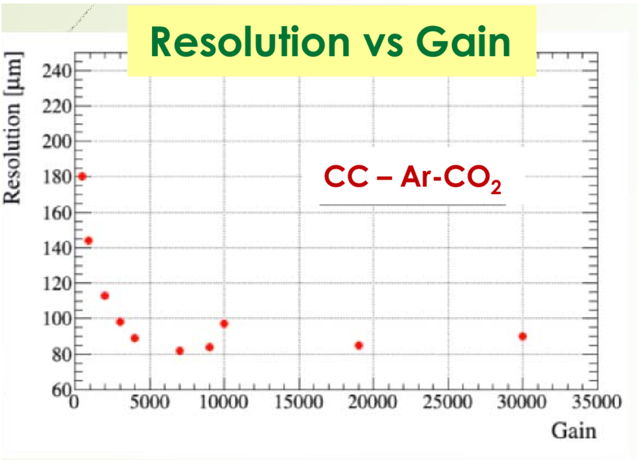


Fig. A.17: Relationship between the gain and the resolution

When one electron is ionized and enlarged by the CGEM, the anode will collect about 6000 electrons, that corresponds to 1 fC, so the down limit output charge is 1 fC. The upper limit output charge is the all the ionized electrons flow into two strip when the charged particle goes through the CGEM with the incident angle zero degree.

Through the simulation count, the dE/dx integral in the drift distance (5 mm), the ionized electrons can be gotten. From the simulation, there are about maximum 28.75 fC electrons for 5 mm ionization gap and 6000 gas gain [48] output to the strip. Considering some charge margin, the maximum 50 fC will be output to one strip. So the charge collected by one strip is from 1 fC to 50 fC.

Last but not least, the power consumption is limited by the space and the cooling capacity. Considering these conditions, every channel cannot exceed 10 mW power consumption.

Based on what discussed above, the detector FEE has the following condition parameters, referring to Table A.2.

Table A.2: Detector condition parameters for the FEE

Input charge	1-50 fC
Detector capacitance	100 pF
Power consumption	10 mW /channel
Event rate	60 kHz

The front-end electronic indicators depend on the detector feature. Based on the detector conditions, the FEE must reach the following indicators in order to meet the CGEM-IT physics demands, which are shown in Table A.3.

Table A.3: FEE demand indicators

Accept charge	1 fC-50 fC
Time resolution (jitter)	<6 ns @1 fC
ENC	< 830 e @100 pF
Integral nonlinearity	< 1% @1 fC-50 fC
Dead time	<1 μs

A.4 Front-End ASIC Design

According to the last section, one FEE board should host 128 channels. The classical discrete components cannot finish so many channels in the limited space, so the ASIC technology is the only choice. Considering the cost and the application in China, the UMC 110 nm CMOS technology is used.

From the perspective of integration, the ASIC chip is better to have more channels. But when one chip integrates too much channel, there will be the power management problem, the heat dissipation problem, and so on. As a compromise, each ASIC chip includes 64 channels, and the power consumption is less than $10\ mW \cdot 64 = 640\ mW$.

Among of the front-end architectures, introduced in Chapter 1, the TOT architecture and the sample-hold architecture can be suitable for this kind of request. It needs good noise and time jitter performance.

The jitter has three main sources. One is the time walk. It happens when the different signal amplitudes cross the same threshold. All the signals have the same shape, so the larger signal has the greater slope and crosses the threshold earlier than the small signal. This time difference is the time walk. Fortunately, the time walk error can be corrected by the calibration.

Another jitter error comes from the detector signal. A typical triple GEM signal is shown in Figure A.18. When the electron clouds of one batch come to the GEM holes, they are not at the same time, so the GEM will output different timing signal even if the whole charge is the same.

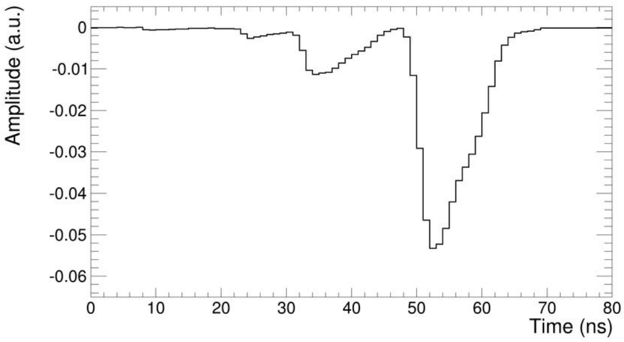


Fig. A.18: Typical triple GEM signal shape

Through the simulation and comparing with other similar detectors, the CGEM outputs the signal as in Figure A.19.

The output don't have the ion tail and have a similar shape, therefore they have a good multi-track and time resolution. The signal duration time is from 30 *ns* to 50 *ns*: 30 *ns* to 40 *ns* the rising time, and 10 *ns* the falling time.

The third jitter source comes from the front-end noise, and it is also influenced by the rising edge slope. An example of the rising edge with noise is shown in Figure A.20. When the rising edges cross the threshold voltage, because of the noise, the time of signal could be earlier or later than the truth signal. So the timing jitter will be generated.

The jitter σ_t caused by the noise can be shown as follows:

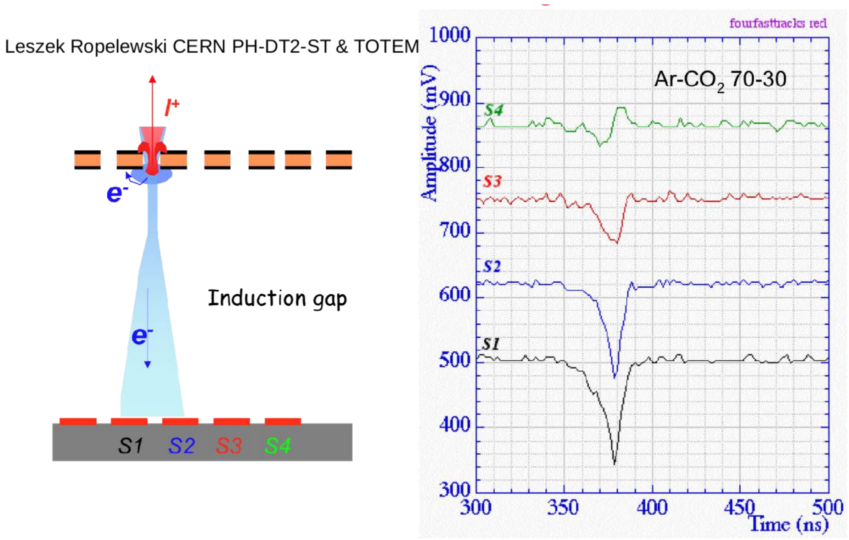


Fig. A.19: A CGEM output signal shape

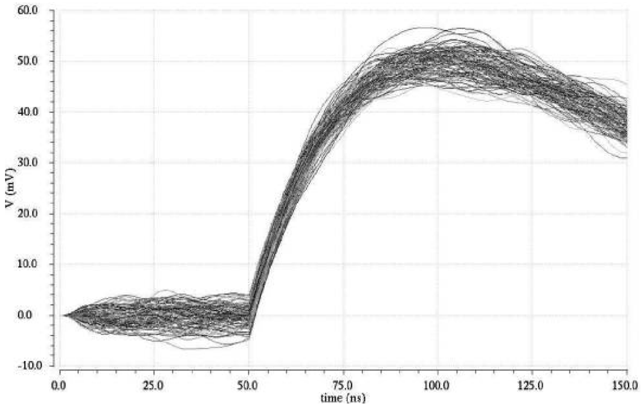


Fig. A.20: Rising edge shape with noise

$$\sigma_t = \frac{\sigma_V}{\left.\frac{dV}{dt}\right|_{t=t_0}} \tag{A.13}$$

where t_0 is the time when the signal crosses the threshold voltage; σ_V is the FEE noise.

The timing jitter is given by the signal noise over the signal slope at the threshold. The FEE bandwidth can be considered as the cut-off frequency of the low-pass filter, so the FEE noise power density spectrum (mainly the thermal noise, also the white

Gaussian noise) will be proportional to the bandwidth. The FEE noise is proportional to the square root of the bandwidth:

$$\sigma_V = A \cdot \sqrt{F_{BW}} \tag{A.14}$$

where A is the thermal noise coefficient determined by the FEE. The FEE bandwidth and the signal rising time have the following relationship:

$$T_{rise} \cdot F_{BW} \approx 0.35 \tag{A.15}$$

So, the FEE noise will be inversely proportional to the square root of the rising time.

$$\sigma_V \approx A \cdot \frac{\sqrt{0.35}}{\sqrt{T_{rise}}} \tag{A.16}$$

The signal slope is directly proportional to the bandwidth:

$$\left.\frac{dV}{dt}\right|_{t=t_0} = B \cdot F_{BW} \tag{A.17}$$

Where B is a constant, together with the equation (A.15), there is:

$$\left.\frac{dV}{dt}\right|_{t=t_0} \approx B \cdot \frac{0.35}{T_{rise}} \tag{A.18}$$

Combining the equation (A.16) and (A.18), the jitter is:

$$\sigma_t = \frac{\sigma_V}{\left.\frac{dV}{dt}\right|_{t=t_0}} \approx \frac{A}{B \cdot 0.35} \cdot \sqrt{T_{rise}} \tag{A.19}$$

This equation means the shorter T_{rise}, the better the jitter, if the rising time is only determined by the FEE. But actually, the detector also contribute the rising time. The whole rising time T_{tot} is:

$$T_{tot} = \sqrt{T_{det}^2 + T_{rise}^2} \tag{A.20}$$

where T_{det} is the detector rising time, which only influences the rising slope and does not influence the FEE noise σ_V. So there is the following result:

$$\sigma_t \propto \frac{\sqrt{T_{rise}^2 + T_{det}^2}}{\sqrt{T_{rise}}} = \sqrt{T_{rise} + \frac{T_{det}^2}{T_{rise}}} \tag{A.21}$$

Differentiating this equation with respect to T_{rise}, the minimum $2 \cdot T_{dec}$ can be gotten when the $T_{rise} = T_{det}$. According to what has been discussed, the FEE should output the timing information with small jitter and the energy information with small noise. So it has the following design:

Firstly, the detector signal flows into one preamplifier. The preamplifier should be a charge sensitive amplifier which has an excellent noise performance. Then the preamplifier splits into two branches: one branch is to detect the energy information, and the other is to detect the timing information. It is shown in Figure A.21.

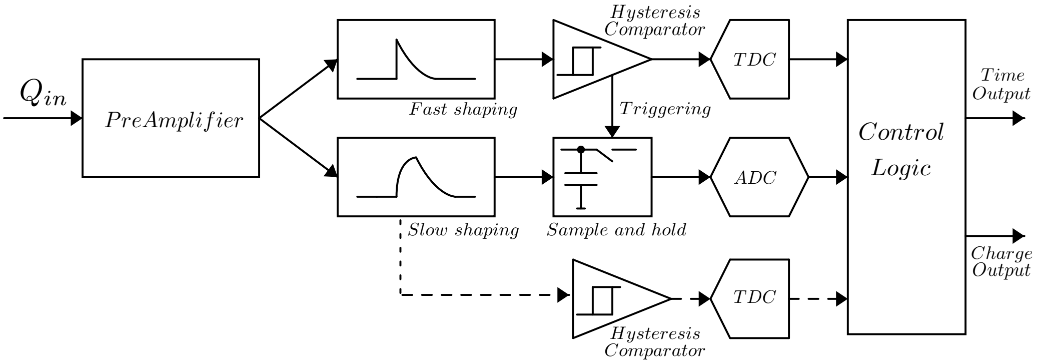


Fig. A.21: FEE architecture with two outputs

The energy branch uses the slow shaping and is sampled by the ADC or the TDC; the timing branch uses the fast shaping and is sampled by the TDC. At last these timing and energy information is output to the control logic.

The timing branch should be designed with a rising time almost the same as the detector rising time (about 40 *ns*), in order to minimize the jitter. That means the bandwidth of the time shaper should be:

$$F_{BWT} = \frac{0.35}{T_{rise}} = \frac{0.35}{40\ ns} = 8.78\ MHz \tag{A.22}$$

where F_{BWT} is the timing branch bandwidth. To not influence the fast signal passing through, the preamplifier bandwidth should be larger than F_{BWT} of 8.78 MHz. The energy branch should have a lower frequency, which can result in the fewer white noise, but the lower limit is determined by the event rate.

To satisfy the 60 kHz, the timing branch and the energy branch signal dead time should be less than 1 μs, and the rising time should be less than 300 ns. So the energy and timing branch lower limited frequency F_{LOW} is:

$$F_{LOW} = \frac{0.35}{T_{rise}} = \frac{0.35}{300\ ns} = 1.17\ MHz \tag{A.23}$$

A.4.1 Preamplifier Design

According to the analysis above, the CSA has been chosen as the preamplifier, which is shown in Figure A.22. The CSA have a negative feedback circuit which consists of the feedback resistor R_f and capacitor C_f. The closed-loop gain is:

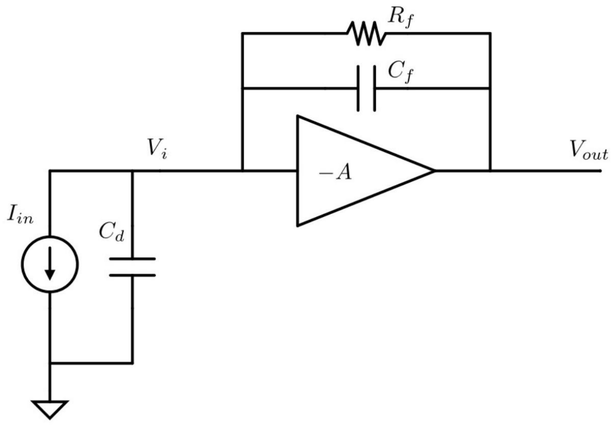


Fig. A.22: A CSA architecture used as the preamplifer

$$Gain_{close} = \frac{1 + sR_f(C_f + C_d)}{1 + sR_fC_f} \tag{A.24}$$

where C_d includes the detector capacitance, the input transistor gate capacitance and other parasitic capacitance between the input terminal and the ground. The $Gain_{close}$ means the closed-loop gain which has one pole and one zero points:

$$Pole : P_1 = \frac{1}{R_f C_f} \tag{A.25}$$

$$Zero : Z_1 = \frac{1}{R_f(C_f + C_d)} \tag{A.26}$$

Normally the amplifier is not the ideal OPA (OPerational Amplifier). When the OPA has one low frequency pole and the crossing lope is $-20dB/dec$, the OPA can be expressed as:

$$A(s) = -A_{OL}\frac{\omega_0}{s + \omega_0} \tag{A.27}$$

where A_{OL} is the amplifier open-loop DC gain; $\omega_0 = 2\pi f_0$, and f_0 is the open-loop cutoff frequency. The open-loop and closed-loop amplitude bode plots are shown in Figure A.23.

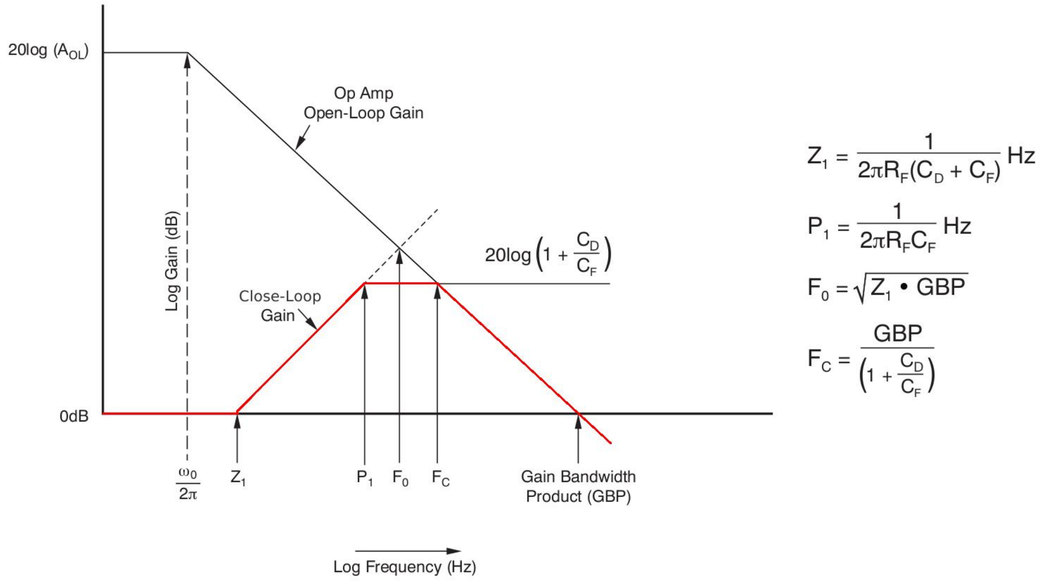


Fig. A.23: CSA architecture bode plot

Between the open-loop gain and the closed-loop gain is the loop-gain. The intersection point of the open-loop and the closed-loop F_C is the CSA bandwidth which influences the rise time. Figure A.23 shows that:

$$F_c = \frac{GBW}{1+\dfrac{C_d}{C_f}} = GBW \cdot \frac{C_f}{C_f + C_d} \tag{A.28}$$

where *GBW* is the Gain Bandwidth Product and also the open-loop system bandwidth.

It can be seen that the feedback capacitor C_F can influence the CSA bandwidth. When C_F is larger, F_C will increase, then the cross point will further from the first pole, and nearer to the second pole of the open-loop system. Therefore, the phase margin will be fewer, and the CSA will become more unstable. Conversely, the fewer C_F makes the bandwidth narrower and the system more stable. So the feedback capacitor should choose the suitable value which can maintain the essential bandwidth. The wider bandwidth can result in losing the phase margin.

In this design, the detector rise time is about 40 *ns*, which corresponds to the frequency:

$$F_{dec} \approx \frac{0.35}{40\ ns} = 8.75\ MHz. \tag{A.29}$$

The detector signal will go through the CSA, and then the timing branch shaper will filter it further. At the end, make the FEE rising time about the same rise time 40 *ns*. So the CSA should have a wider bandwidth, which makes most of the detector signal (less than 8.75 *MHz*) go through without loss. Considering some bandwidth margin, it chooses the CSA rising time 20 *ns*, which corresponds to the frequency 17.5 *MHz*.

The bandwidth of the CSA is limited by the intrinsic frequency $\mathrm{f_T}$. The relationship between the intrinsic frequency and the channel length *L* is shown in Figure A.24[2].

The reference and figure shows that in CMOS processes with channel lengths below 130 *nm*, the $\mathrm{f_T}$ can easily exceed 100 *GHz*. So in the 110 *nm* technology, the $\mathrm{f_T}$ is beyond 100 *GHz*. it is achievable to set the CSA open-loop bandwidth 10 *GHz*. In this situation, by the equation (A.28), the feedback capacitor can be counted:

$$C_f = \frac{C_d}{\dfrac{GBW}{F_c} - 1} = \frac{100\ pF}{\dfrac{10\ GHz}{17.5\ MHz} - 1} = 175\ fF \tag{A.30}$$

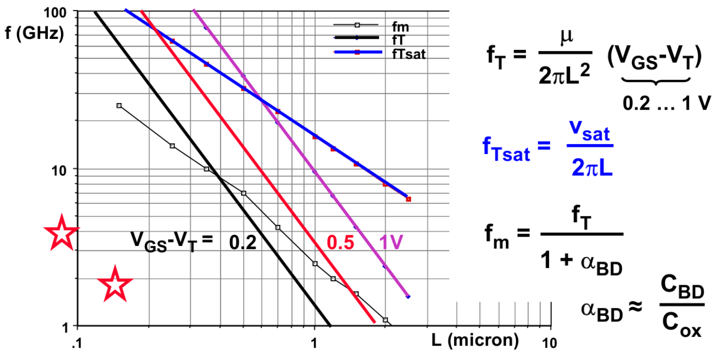


Fig. A.24: The maximum intrinsic frequency f_T versus the channel length L[2]

To get a better phase margin, set the feedback capacitor 150 fC. The MOM capacitor is selected as the feedback capacitor which has the better matching.

According to the Miller theorem, the equivalent input capacitor of the feedback capacitor is:

$$C_{equi} = C_f \cdot A_{OL} \tag{A.31}$$

The equivalent capacitance and the input capacitance will receive the detector charge together. In order to reduce the C_d influence, make the most of the charge flow into the feedback capacitor. It should be:

$$C_d \leq \frac{1}{100} \cdot (C_d + C_{equi}) = \frac{1}{100} \cdot (C_d + C_f \cdot A_{OL}) \tag{A.32}$$

$$A_{OL} \geq \frac{100 \cdot C_d - C_d}{C_f} = \frac{99 \cdot C_d}{C_f} = 66000 \tag{A.33}$$

So the open-loop gain of the CSA should be larger than the 66000 with the bandwidth 10 GHz.

Based on Figure A.22, when the detector current i_i flows into u_i node, according to the KCL (Kirchhoff's Circuit Laws) and the amplifier feature, there are:

$$i_{\mathrm{i}} = \frac{v_{\mathrm{i}}}{\frac{1}{sC_{\mathrm{d}}}} + \frac{v_{\mathrm{i}} - v_{\mathrm{o}}}{\frac{1}{\frac{1}{R_f} + sC_f}} \tag{A.34}$$

$$\frac{v_{\mathrm{o}}}{v_{\mathrm{i}}} = -A(s) = -A_0 \frac{-s_0}{s - s_0} = -A_0 \frac{\omega_0}{s + \omega_0} \tag{A.35}$$

From the above two equations, there is:

$$\frac{v_\mathrm{o}}{v_\mathrm{i}} = -AR_f \frac{1}{sC_dR_f + (1+A)(1+sC_fR_f)} \tag{A.36}$$

When A is a function:

$$A = A_{OL}\frac{\omega_0}{s+\omega_0} \tag{A.37}$$

There is:

$$\frac{v_o}{i_\mathrm{i}} = \frac{-A_{OL}R_f\omega_0}{s^2(C_d+C_f)R_f + s(1+\omega_0C_dR_f + \omega_0(1+A_{OL})C_fR_f) + \omega_0(1+A_{OL})} \tag{A.38}$$

Normally, the A_{OL} is far greater than one, so this equation can be simplified to:

$$\frac{v_o}{i_\mathrm{i}} \approx \frac{-A_{OL}R_f\omega_0}{s^2(C_d+C_f)R_f + s(1+\omega_0C_dR_f + \omega_0A_{OL}C_fR_f) + \omega_0A_{OL}} \tag{A.39}$$

$$\frac{v_o}{i_\mathrm{i}} = T(s) \approx \frac{-R_f}{s^2\frac{(C_d+C_f)R_f}{A_{OL}\omega_0} + s\frac{1+\omega_0R_f(C_d+C_fA_{OL})}{A_{OL}\omega_0} + 1} \tag{A.40}$$

Rewrite the denominator into the following form, here suppose it has two widely spaced real poles:

$$D(s) = (s\tau_f+1)(s\tau_r+1) = s^2\tau_f\tau_r + s(\tau_f+\tau_r) + 1 \approx s^2\tau_f\tau_r + s\tau_f + 1 \tag{A.41}$$

where suppose the τ_f is much greater than the τ_r. Comparing the two formulas, there is:

$$\tau_f = \frac{1+\omega_0R_f(C_d+C_fA_{OL})}{A_{OL}\omega_0} \tag{A.42}$$

$$\tau_f\tau_r = \frac{(C_d+C_f)R_f}{A_{OL}\omega_0} \tag{A.43}$$

Because the A_{OL} is greater than one, and $C_f A_{OL}$ is greater than C_d, there is:

$$\tau_f \approx R_f C_f \tag{A.44}$$

τ_r is:

$$\begin{aligned}\tau_r = \frac{\tau_r \tau_f}{\tau_f} = \frac{\dfrac{(C_d + C_f)R_f}{A_{OL}\omega_0}}{\dfrac{1 + \omega_0 R_f (C_d + C_f A_{OL})}{A_{OL}\omega_0}} &= \frac{(C_d + C_f)R_f}{1 + \omega_0 R_f (C_d + C_f A_{OL})} \\ &\approx \frac{1}{\omega_0 A_{OL} \dfrac{C_f}{C_f + C_d}}\end{aligned} \tag{A.45}$$

where $\omega_0 A_{OL} C_f/(C_f + C_d)$ is the closed-loop bandwidth F_c multiply the 2π. In this design, the open-loop bandwidth is set about 10 GHz, and the feedback capacitor C_f and detector capacitance C_d are 150 fF and 100 pF respectively. The τ_r is:

$$\tau_r \approx \frac{1}{10\ GHz \cdot 2\pi \cdot \dfrac{150\ fF}{150\ fF + 100\ pF}} = 10.63\ ns \tag{A.46}$$

If the feedback resistor is infinity, the transfer function $T(s)$ is:

$$T(s) = \frac{R_f}{(1 + s\tau_f)(1 + \tau_r)} = \frac{R_f}{(1 + sR_f C_f)(1 + \tau_r)} = \frac{1}{sC_f} \cdot \frac{1}{1 + s\tau_r} \tag{A.47}$$

When the Dirac-delta with the total charge Q_{in} input, using the Inverse Laplace Transform, the output voltage is:

$$V_{out}(t) = \frac{Q_{in}}{C_f}\left(1 - e^{-\frac{t}{\tau_r}}\right) \tag{A.48}$$

The output voltage will rise up according to one constant τ_r, and the rising time is about $2.2\tau_r$. Then the output voltage will be Q_{in}/C_f at the end.

If the rising constant τ_r can be ignored, the transfer function $T(s)$ is:

$$T(s) = \frac{R_f}{1 + sR_fC_f} \tag{A.49}$$

When a Dirac-delta with the total charge Q_{in} is the input, using the Inverse Laplace Transform, the output voltage is:

$$V_{out}(t) = \frac{Q_{in}}{C_f} e^{-\frac{t}{\tau_f}} \tag{A.50}$$

The output voltage will become Q_{in}/C_f immediately, and then the voltage will recover to zero as a constant τ_f.

If the feedback resistor R_f is limited, and τ_r is not zero, when a Dirac-delta with the total charge Q_{in} is the input, using the Inverse Laplace Transform, the output voltage is:

$$V_{out}(t) = Q_{in} \frac{R_f}{\tau_r - \tau_f} \left(e^{-\frac{t}{\tau_r}} - e^{-\frac{t}{\tau_f}} \right) = \frac{Q_{in}}{C_f} \frac{\tau_f}{\tau_r - \tau_f} \left(e^{-\frac{t}{\tau_r}} - e^{-\frac{t}{\tau_f}} \right) \tag{A.51}$$

Differentiating the above equation. It can get the peaking time T_p, at which the output voltage V_{out} will reaches the maximum value:

$$\frac{dV_{out}}{dt} = 0 \rightarrow \frac{1}{\tau_f} e^{-\frac{T_p}{\tau_f}} = \frac{1}{\tau_r} e^{-\frac{T_p}{\tau_r}} \rightarrow T_p = \frac{\tau_f \tau_r}{\tau_r - \tau_f} \ln \frac{\tau_r}{\tau_f} \tag{A.52}$$

Because of the limited feedback resistor, the peaking time T_p is larger than the $2.2\tau_r$. Substitute T_p into the first term of the bracket in the equation (A.56).

$$e^{-\frac{T_p}{\tau_r}} = e^{-\frac{1}{\tau_r} \frac{\tau_f \tau_r}{\tau_r - \tau_f} \ln \frac{\tau_r}{\tau_f}} = \left(e^{\ln \frac{\tau_f}{\tau_r}} \right)^{\frac{\tau_f}{\tau_r - \tau_f}} = \left(\frac{\tau_f}{\tau_r} \right)^{\frac{\tau_f}{\tau_r - \tau_f}} \tag{A.53}$$

Similarly, the second item can be expressed as:

$$e^{-\frac{T_p}{\tau_f}} = \left(\frac{\tau_f}{\tau_r} \right)^{\frac{\tau_r}{\tau_r - \tau_f}} \tag{A.54}$$

So the bracket of the equation (A.56) can be expressed as:

$$\begin{aligned} e^{-\frac{T_p}{\tau_r}} - e^{-\frac{T_p}{\tau_f}} &= \left(\frac{\tau_f}{\tau_r}\right)^{\frac{\tau_f}{\tau_r-\tau_f}} - \left(\frac{\tau_f}{\tau_r}\right)^{\frac{\tau_r}{\tau_r-\tau_f}} = \left(\frac{\tau_f}{\tau_r}\right)^{\frac{\tau_r}{\tau_r-\tau_f}} \left[\left(\frac{\tau_f}{\tau_r}\right)^{\frac{\tau_f-\tau_r}{\tau_r-\tau_f}} - 1\right] \\ &= \left(\frac{\tau_f}{\tau_r}\right)^{\frac{\tau_r}{\tau_r-\tau_f}} \left(\frac{\tau_r}{\tau_f} - 1\right) = \frac{\tau_r - \tau_f}{\tau_f} \left(\frac{\tau_f}{\tau_r}\right)^{\frac{\tau_r}{\tau_r-\tau_f}} \end{aligned} \tag{A.55}$$

Put the equation (A.55) into the equation (A.56), there is:

$$V_{out}(T_p) = \frac{Q_{in}}{C_f} \left(\frac{\tau_f}{\tau_r}\right)^{\frac{\tau_r}{\tau_r-\tau_f}} = \frac{Q_{in}}{C_f} \left(\frac{\tau_f}{\tau_r}\right)^{\frac{1}{1-\frac{\tau_f}{\tau_r}}} \tag{A.56}$$

From the equation, it can be seen that the larger τ_f/τ_r makes the less difference between the ideal value Q_{in}/C_f and the actual value. The difference value has one special name "Ballistic Loss". The Ballistic loss can influence the linear degree, and further influence the energy resolution. In this design, to reduce the Ballistic loss to less than 5%, there is:

$$\frac{\tau_f}{\tau_r} \geq 100 \rightarrow \tau_f \geq 100\tau_r \rightarrow R_f C_f \geq 100\tau_r \tag{A.57}$$

According to the above analysis, when $\tau_r = 10.63\ ns$, the feedback capacitor C_f is 150 fF. So R_f is:

$$R_f \geq 100 \cdot \frac{\tau_r}{C_f} = 100 \cdot \frac{10.63\ ns}{150\ fF} = 7\ M\Omega \tag{A.58}$$

Such large resistance cannot be realized by a real resistor in the COMS technology. Because not only the large resistor occupies large silicon area, but also such large resistor will generator large parasitic capacitance, which can make the system unstable.

In this design, the large resistor is realized by transistors, as shown in Figure A.25. The equivalent feedback resistor consists of the M_0, M_1, M_2. The equivalent feedback resistance value comes from voltage difference variation divided by current variation in the two terminals. The equivalent resistance is calculated as follows.

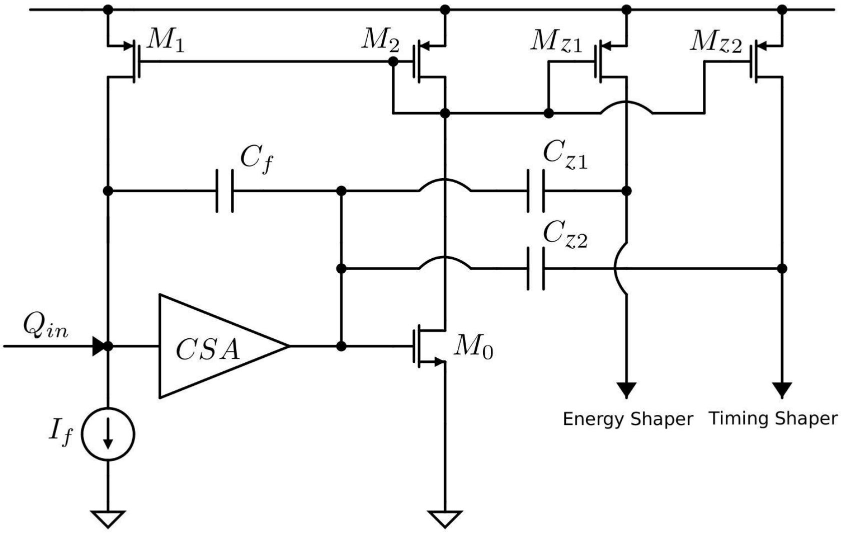


Fig. A.25: Feedback capacitor and equivalent resistor (M_0, M_1, M_2) of the Charge Sensitive Amplifier

When the CSA output V_o changes of δV_o, this voltage will be changed into the current by M_0:

$$\delta I_0 = \delta V_o \cdot g_{m0} \tag{A.59}$$

The δI_0 will be transferred onto the $M1/M2$ gate voltage and then be changed into δV_{M2g} by the diode connection M_2:

$$\delta V_{M2g} = \delta I_0 \cdot \frac{1}{g_{m2}} \tag{A.60}$$

$M1$ changes δV_{M2g} into the feedback current δI_i:

$$\delta I_i = \delta V_{M2g} \cdot g_{m1} \tag{A.61}$$

Combined the equation (A.59), (A.60) and (A.63), there are:

$$\delta I_i = \delta V_o \cdot g_{m0} \cdot \frac{1}{g_{m2}} \cdot g_{m1} \tag{A.62}$$

$$\frac{\delta V_o}{\delta I_i} = R_f = \frac{g_{m2}}{g_{m1} \cdot g_{m0}} \tag{A.63}$$

Setting $M0, M1, M2$ transconductance as 9.2 μS, 191 nS and 19 μS respectively, R_f turns to be:

$$R_f = \frac{g_{m2}}{g_{m1} \cdot g_{m0}} = \frac{19\ \mu S}{191\ nS \cdot 9.2\ \mu S} = 10.8\ M\Omega \tag{A.64}$$

From this equation, R_f is larger than 7 $M\Omega$, which fulfils the requirement. The current I_f is used for biasing $M1$ and setting the CSA output baseline voltage V_{base}. The capacitors C_{z1} and C_{z2} are used for the differential function. The M_0, M_2, M_{z1} constitutes the equivalent resistance R_{z1}; and M_0, M_2, M_{z2} constitutes the equivalent resistance R_{z2}. The R_{z1} and R_{z2} are used as the pole-zero cancellation. Using a similar method of calculating the feedback resistance, it can set the transconductance of M_{z1} and M_{z2} as the same 3.87 μS, and get R_{z1} and R_{z2} as about one twentieth the feedback R_f is 533 $k\Omega$.

The intrinsic gain of the 110 nm technology is about 100, so to get the open-loop gain of the CSA more than 66000, the cascode architecture and at least three levels cascades must be used. Based on this situation, the architecture in Figure A.26 is used.

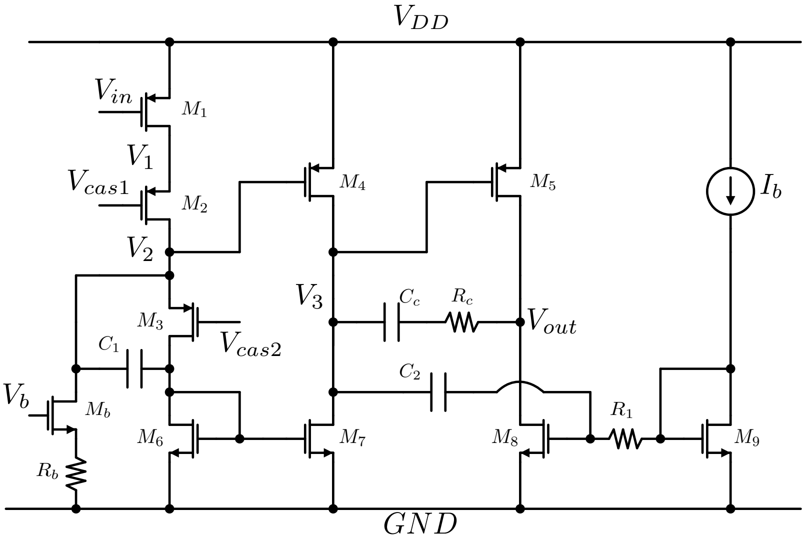


Fig. A.26: Transistor level design of the CSA

In the CSA architecture, the most important transistor is the input transistor. Because the input transistor contributes normally about 50-70% ENC noise. In order to minimize the noise, the transconductance must be enlarged. The current effect (transconductance over current) versus the inversion coefficient is shown in Figure A.27.

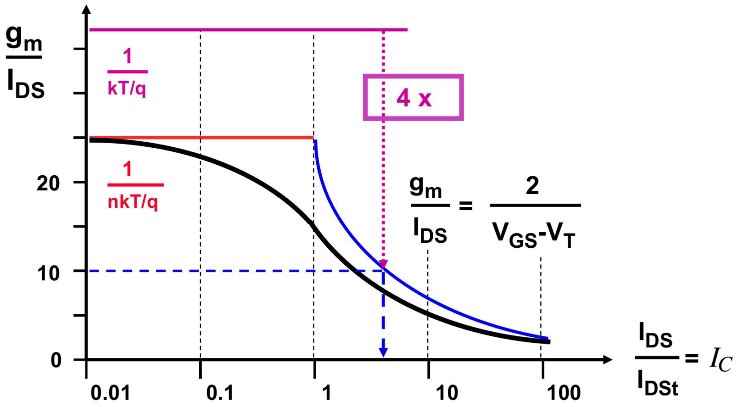


Fig. A.27: Current effect versus the inversion coefficient[2]

It can be seen that in order to get a better current effect, to reduce the power consumption, the lower inversion coefficient is necessary. When the $I_C = 0.1$, gm/I_{ds} has a suitable margin effect. So here set $I_C = 0.1$, and get g_m/I_{ds} about 23 and γ=0.515.

The thermal noise of the input transistor will be the main contribution source. The flick noise is proportional to the detector capacitance and has no relationship with the signal peaking time. The detector capacitance is small and the flick noise can reduce by enlarging the transistor width and length at the same proportion. Comparing the thermal noise, the flick noise can be ignore. The thermal noise in the FEE can be expressed [1]:

$$ENC_i = \frac{1}{q} v_{nw} C_T \sqrt{N_w \frac{1}{T_p}} \tag{A.65}$$

The T_p is limited by the pulse width. The dead time, which has been discussed, is less than 1 μs. Considering some margin, here set the pulse width is about 800 ns. The relationship between T_p and the dead time is shown in Figure A.28 [1]. This figure shows that when using the real poles, the ratio of order 3 is about 5. That means the peaking time should be 800 $ns/5$ = 160 ns.

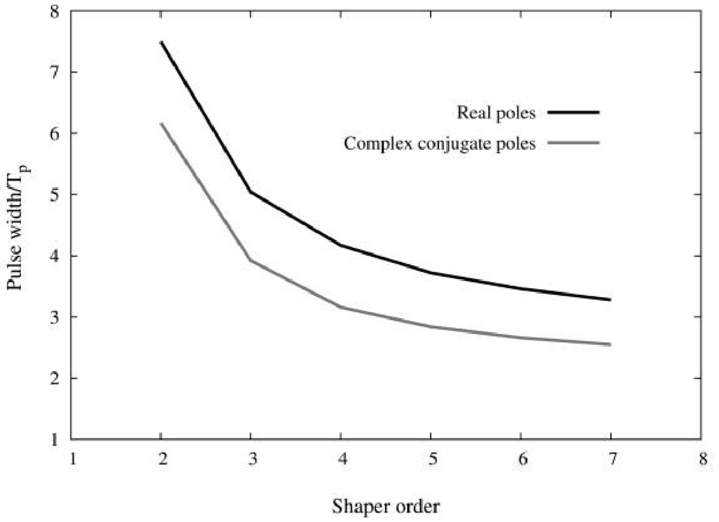


Fig. A.28: Ratio between the pulse width and the peaking time for the shapers with different orders

When the rise time T_p is 160 *ns*, $C_T = 100\ pF$, ENC_i is about 50% *ENC*, that is about:

$$ENC_i = 50\% \cdot ENC = \sqrt{770\ e \cdot 770\ e \cdot 50\%} = 540\ e \tag{A.66}$$

The N_w is about 1, which can be checked in the book [1]. The equation (A.65) can be changed into:

$$g_m = \frac{4k_B T \gamma C_T^2}{ENC_i^2}\frac{N_w}{T_p} = \frac{4 \cdot 1.38 \cdot 10^{23}\ J/K \cdot 300\ K \cdot 0.515}{(540 \cdot 1.6^{-19}\ C)^2} \cdot \frac{(100\ pF)^2}{160\ ns} = 69\ mS \tag{A.67}$$

So the input transistor should have 69 *mS* at least. When the current effect is about 23, the input current needs at least 3 *mA*, which can consume about 3.6 *mW*. The whole FEE power consumption can be controlled within 5 *mW*.

Through the simulation, the CSA open-loop gain curve is shown in Figure A.29. The open-loop gain can reach 10^5 which fulfils the requirement of the lower limit 7×10^4. The bandwidth is about 2 *GHz*, but the cross point slope is about -40 *dB/dec*. The parameters are counted according to the -20 *dB/dec*, so drawing an extension line from the -20 *dB/dec*. The cross point is 15 *GHz*, which fulfils the requirement of 10 *GHz*.

From what discussed above, the CSA preamplifier can fulfil the requirements of the parameters.

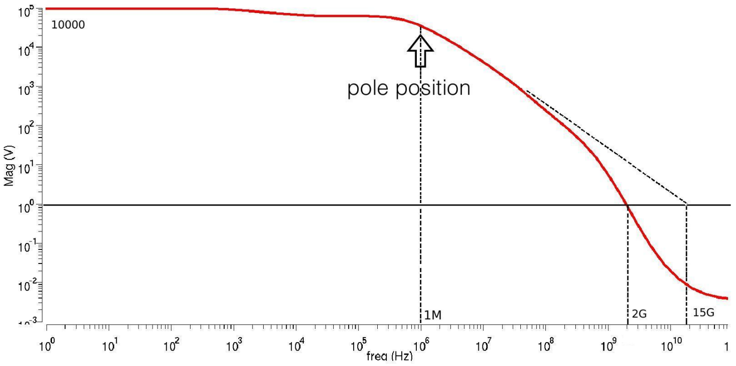


Fig. A.29: Open-loop gain and bandwidth of the CSA

The main pole is on the arrow point, about at 1 MHz. In Figure A.26, the main pole is caused by the V_3 node, where it has a large impedance r_{DS} and large Miller capacitance. The width ratio between M_6 and M_7 is also very important to the rise time. Enlarging the ratio, then the M_4, M_7 current increase. The V_3 impedance is the parallel with the M_4, M_7 output resistance which is $1/(\lambda \cdot I_d)$. So when the V_3 impedance decreases, the main pole increases. At the same time, the open-loop gain depending on the ratio also increases. The bandwidth increases as well. The closed-loop gain curve (the reciprocal of the feedback transfer function) will cross the open-loop gain curve further than the original point. So it will be nearer to the second open-loop pole. Vice versa, the phase margin will fewer, at the same time it will also consume more power. Here there is a trade off: higher ratio, higher power consumption, faster rise time, more unstable, or the contrary.

A.4.2 Shaper Design

Following the CSA, there are two shapers: the timing branch and the energy branch. The two branches get the current from the CSA. The timing branch will shape or filter the CSA output signal with the 40 ns rise time. The energy branch will shape or filter the output CSA signal with 160 ns rise time. The timing branch with the CSA is shown in Figure A.30.

The C_z is used to differentiate this CSA output and it changes the voltage signal into a current signal. The R_z is used as the pole-zero cancellation. The R_1C_1 integrate this current signal into a voltage signal, and R_c changes the voltage into a current

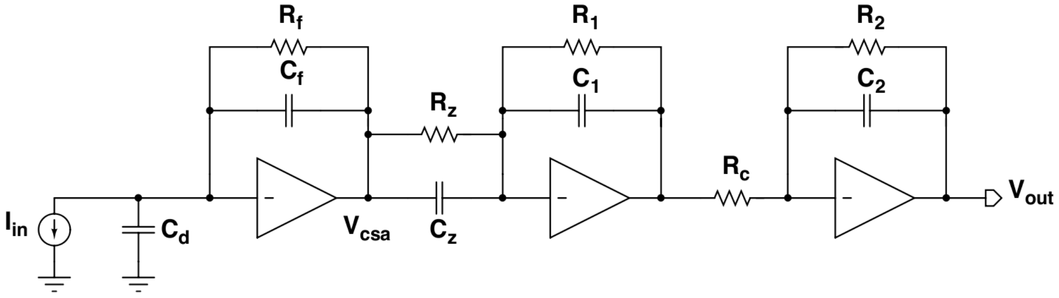


Fig. A.30: Timing branch schematic with the CSA

again. At last R_2C_2 integrates the current signal the second time and changes it into a voltage signal. The transfer function is:

$$V_{out}(s) = I_{in}(s)\frac{R_f}{1+sR_fC_f}\frac{1+sC_zR_z}{R_z}\frac{R_1}{1+sR_1C_1}\frac{1}{R_c}\frac{R_2}{1+sR_2C_2} \tag{A.68}$$

In last section, C_zR_z is set equal to R_fC_f, so the equation can be simple:

$$\frac{V_{out}(s)}{I_{in}(s)} = T(s) = \frac{R_f}{R_z}\frac{R_1R_2}{R_c(1+sR_1C_1)(1+sR_2C_2)} \tag{A.69}$$

This transfer function unit is *ohm*, and set $R_1C_1 = \tau_1$, $R_2C_2 = \tau_2$. It can be gotten that the transfer function has two poles $1/\tau_1$ and $1/\tau_2$. Like the CSA transfer function analysis, when a Dirac-delta conveying the total charge Q_{in} input, the peaking time is:

$$T_{p,Time} = \frac{\tau_1\tau_2}{\tau_2-\tau_1}\ln\frac{\tau_2}{\tau_1} \tag{A.70}$$

The peak voltage is:

$$V_{out,peakT} = \frac{Q_{in}}{C_1}\frac{R_f}{R_z}\frac{R_2}{R_c}\left(\frac{\tau_1}{\tau_2}\right)^{\frac{\tau_2}{\tau_2-\tau_1}} \tag{A.71}$$

Setting the following parameters:

$$R_1 = 100\ k\Omega, R_2 = 20\ k\Omega, C_1 = C_2 = 1\ pF, R_c = 20\ k\Omega, Q_{in} = 1\ fC \tag{A.72}$$

where $\tau_1 = 100\ ns$, $\tau_2 = 20\ ns$, and $R_f = 20R_z$, which has been gotten in the last section. Then $V_{out,peakT} = 13.37\ mV$, which is almost the same as the simulation value 12.8 mV/fC. The peaking time is about $V_{out,peakT} = 40\ ns$.

When the detector rise time is 40 ns, the CSA rise time is 23 ns, and the shaper rise time is 40 ns, the final rise time is $\sqrt{(40\ ns)^2 + (23\ ns)^2 + (40\ ns)^2} = 61\ ns$, which is almost the same as the simulation rise time of 70 ns. In this situation, the jitter of the timing branch is 4.62 ns, which is less than the request value of 6 ns.

The energy branch does not need a the short peaking time, but the contrary needs a longer peaking time within a limited value in order to collect most of the charge. So the complex conjugate poles are more suitable for the energy branch shaper. The architecture is shown in Figure A.31.

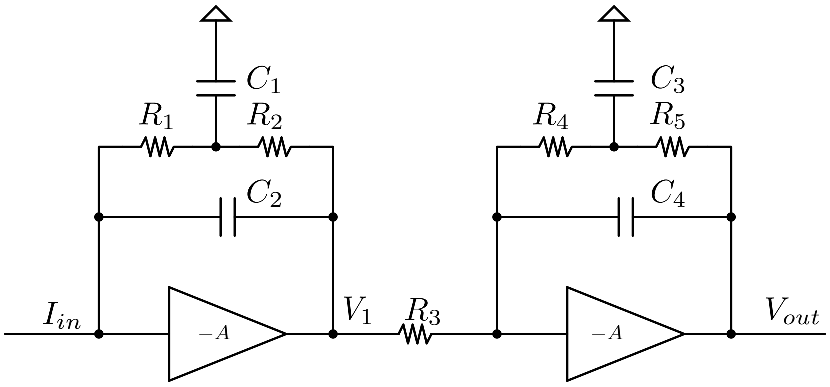


Fig. A.31: Energy branch shaper schematic

Using the similar analysis method as the time shaper, the ENC of the energy shaper is 650 e, and the gain is 12.2 mV/fC. The peaking time is about 200 ns.

The last important part of the shaper is the baseline holder circuit, which can keep the baseline stable at one desired voltage. The baseline circuit can be seen as one high pass filter. The low-frequency variation can be filtered by the baseline holder circuit, which is shown in the full front-end schematic in Figure A.32.

This figure shows the timing branch with baseline holder, and the energy branch is similar with it. In the baseline holder, M_0 can be seen as a current source, which plays two roles. One role is to set the static current source of M_z, and the other is to set the static voltage point at the output of A_1.

When the output of A_1 is set 1.1 V, according to the difference between the 1.1 V and the input voltage of A_1, R_{sh} can be counted. Then C_{sh} can be gotten by τ_{sh}.

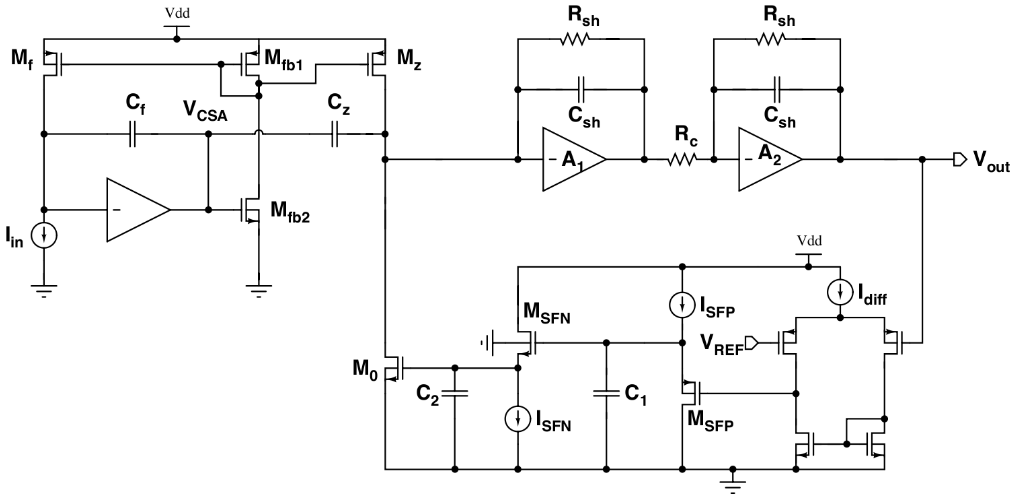


Fig. A.32: Full front-end chain with the baseline holder

The shaper is a high-pass filter and the baseline holder is a low-pass filter, so they together make up a bandpass filter. The baseline holder main pole P_{main} is at the gate of M_0, where has a large capacitance C_2, which corresponds to the time constant τ_{p1}. The second pole P_{sec} node is at the gate of M_{SFN}, which corresponds to the time constant τ_{p2}.

There is a negative feedback between the baseline holder and the shaper. When the output voltage rises, the inverse signal will output at the gate of M_{SFP}. This signal goes through two voltage follower and then goes to the gate of M_0. At this point, because of the large capacitance C_2, the high-frequency signal will go to the ground through C_2. The low-frequency signal such as the baseline variation will change M_0 current. Because of the M_z current stable, the reduced M_0 current reduces the output of A_1. Then further it makes R_c current fewer, which decrease the output low-frequency signal.

The system stable (the baseline holder circuit and the shaper forming a negative feedback system) depends on the loop phase margin. The open-loop gain from the output of M_z to the output is:

$$gain_{open,resist} = \frac{R_{sh} \cdot R_{sh}}{R_c} \cdot \frac{1}{(1+s\tau_{sh})^2} \tag{A.73}$$

Here the unit is Ω. The feedback also changes the output voltage into the input current, the unit is Ω. The open-loop resistance gain is about R_{sh}^2/R_c, which is very large. When a little current varies at the output of the M_z, the output low frequency voltage will change a lot. So the baseline holder is introduced to improve this situation. When the open-loop resistance gain is large, the closed-loop gain depends on the feedback circuit.

$$gain_{feedback,resist} = A_d \cdot g_{m0} \cdot \frac{1}{(1+s\tau_{p1})(1+s\tau_{p2})} \tag{A.74}$$

The closed-loop resistance gain is:

$$\begin{aligned} gain_{closed,resist} &= \frac{gain_{open,resist}}{1+gain_{open,resist} \cdot gain_{feedback,resist}} \approx \frac{1}{gain_{feedback,resist}} \\ &= \frac{(1+s\tau_{p1})(1+s\tau_{p2})}{A_d \cdot g_{m0}} \end{aligned} \tag{A.75}$$

A_d is the differential amplifier gain, and g_{m0} is the transconductance of M_0. The amplitude bode plot is shown in Figure A.33.

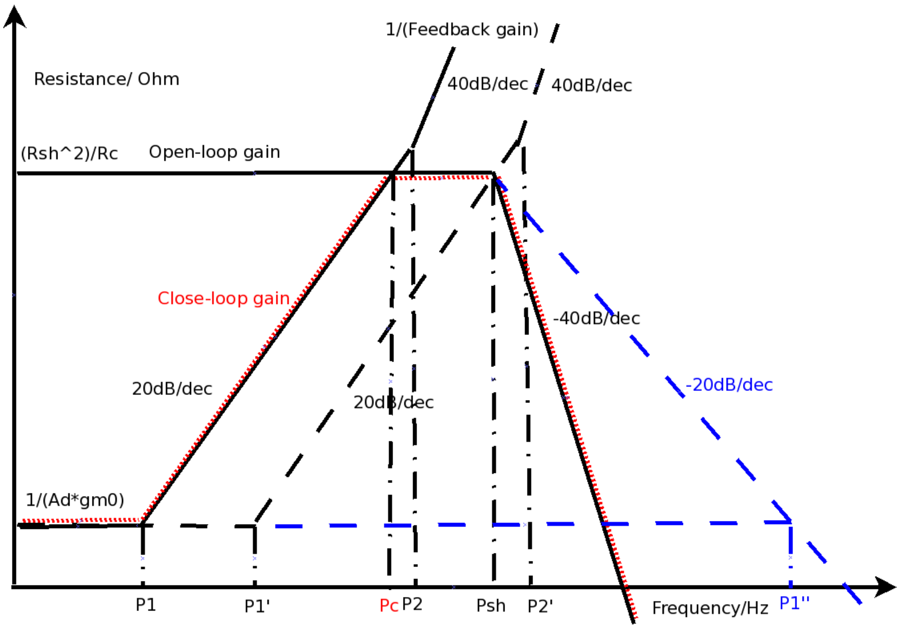


Fig. A.33: Amplitude bode plot of the baseline

In Figure A.33, the loop gain is the open-loop gain minus the feedback gain inverse. The feedback gain line crosses the open-loop gain at the point, which is the unit gain of the loop gain, P_C. To be sure of the unit gain, there is -20 dB/dec. P_C must be smaller than P_{sh}. The dotted line of the 1/feedback gain is the limit of the system stable. In this situation, the second pole P_2' must be larger than P_{sh}. The limit P_1' is:

$$P_1' = \frac{P_{sh}}{\frac{R_{sh}^2/R_c}{1/A_d \cdot g_{m0}}} \tag{A.76}$$

Because of the slope of the 1/feedback gain as 20 dB/dec, when the amplitude reduces X times, the frequency should reduce X time also in the same line. It can draw one -20 dB/dec auxiliary blue line. The dotted black line will be symmetric with the blue auxiliary line with P_{sh} as the symmetry axis. Because of the same bandwidth gain product, when the gain lower X times, the bandwidth larger X times. So P_1'' is X times P_{sh}, at the same time, P_1' is less X times than P_{sh}.

When the low frequency pole moves from P_1' to P_1, the bandpass frequency moves from P_C to P_{sh}. The second pole must be larger than P_C. When the 1/feedback gain is more than P_C, the loop gain becomes less than one. The closed-loop gain will almost the same as the open-loop gain. So the red dotted line will be the same as the open-loop gain. When the main pole is larger than P_1', not only the baseline has some low-frequency oscillation, but also it reduces the output peak value.

Setting $C_2 = 5.2\ pF$ can be realized by the Ncap device. The Ncap device has a large capacitance density. The pole P_1 is:

$$P_1 = \frac{1}{2\pi \cdot C_2 \cdot 1/g_{m,sfn}} = \frac{1}{2\pi \cdot 5.2\ pF \cdot 1/5.78\ nS} = \frac{1100\ rad/s}{2\pi} = 177\ Hz. \tag{A.77}$$

The low frequency gain is:

$$gain_{low} = \frac{1}{A_d \cdot g_{m0}} = \frac{1}{(60.6\ \mu S \cdot (429\ k\Omega//1.5\ M\Omega)) \cdot 41.75\ \mu S} = 1.187\ k\Omega \tag{A.78}$$

The open-loop gain is $R_1 \cdot R_2/R_c = 100\ k\Omega$. The P_C is:

$$P_C = P_1 \cdot \frac{100\ k\Omega}{1.187\ k\Omega} = 9860\ Hz. \tag{A.79}$$

because of the pole P_1, the shift phase at the P_C is:

$$\phi = -tan^{-1}\left(\frac{P_C}{P_1}\right) = -89^\circ \tag{A.80}$$

Using the similar method of P_1, the second pole P_2 is far greater than P_1. So it will not influence the stable. In this design, the baseline voltage is set 300 mV. The baseline holder can control the output baseline voltage well.