

# Communication Systems of Dawning6000

---

## Design and Optimization

HUO Zhigang  
Institute of Computing Technology,  
Chinese Academy of Sciences  
Oct 27, 2010



# Outlines

---

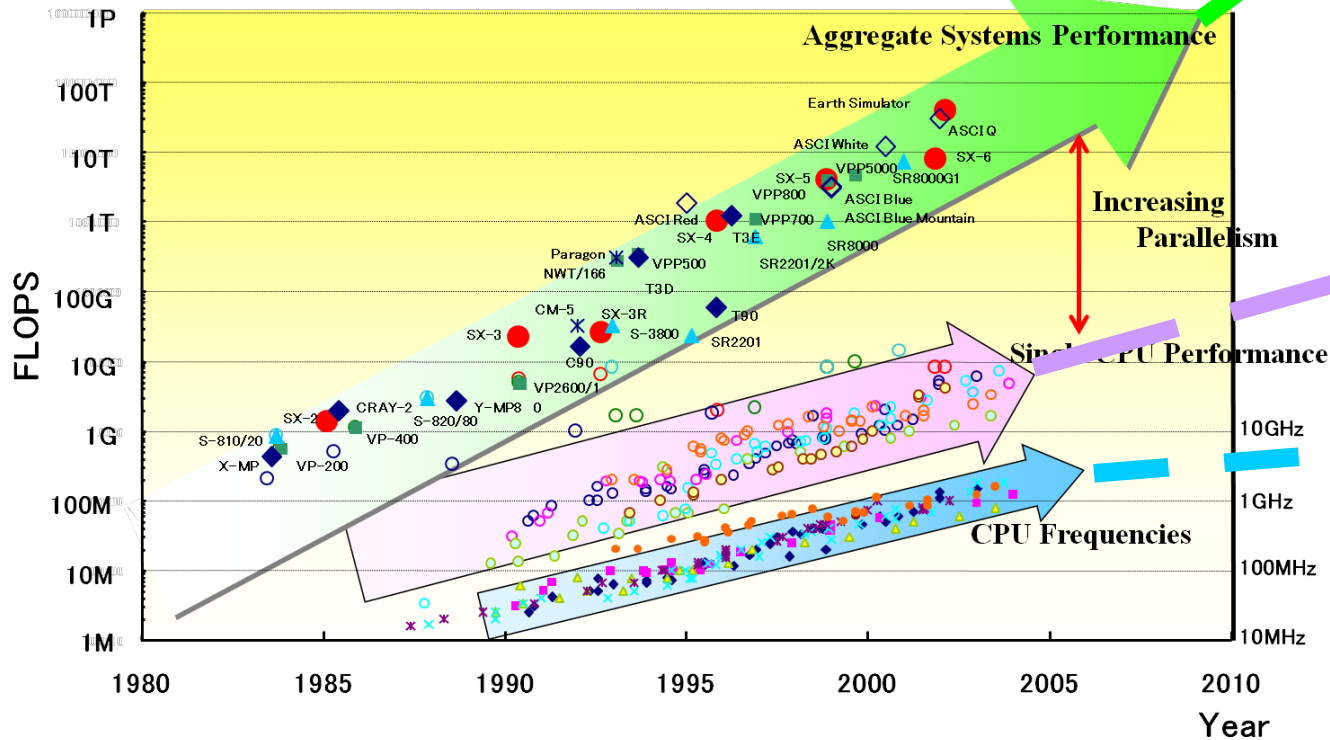
- Introduction of Dawning 6000
- Communication System of X86 Partition
- Communication System of Loongson Partition
  - Experimental Results
- Conclusions
- Acknowledgments



# HPC Development

- Trends of HPC Performance Evolution
  - Where is the limit?

**Power Wall**



Trends of HPC Perf. Increase  
( 1980 - 2010 )



# A Tale of Two Partitions

- Loongson
- Low Price
- Low Power

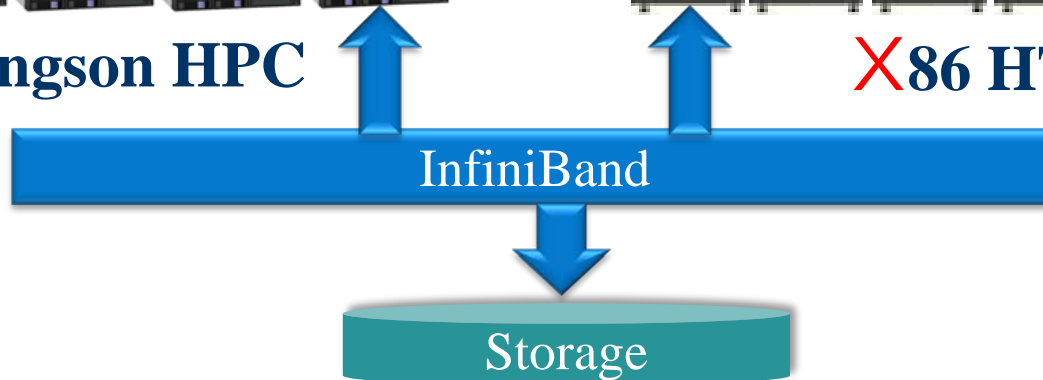
- Xeon
- eXpensive
- Xcelerator



Loongson HPC



X86 HPC



---

Communication System  
in  
the  $X^3$  Partition

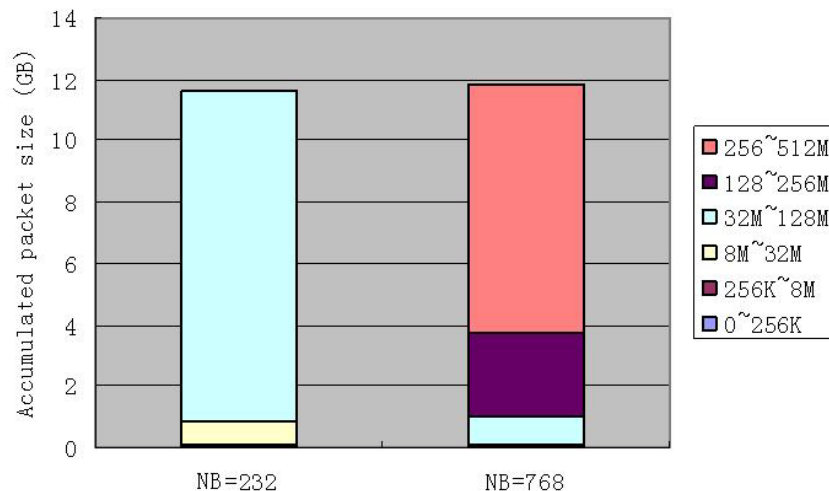
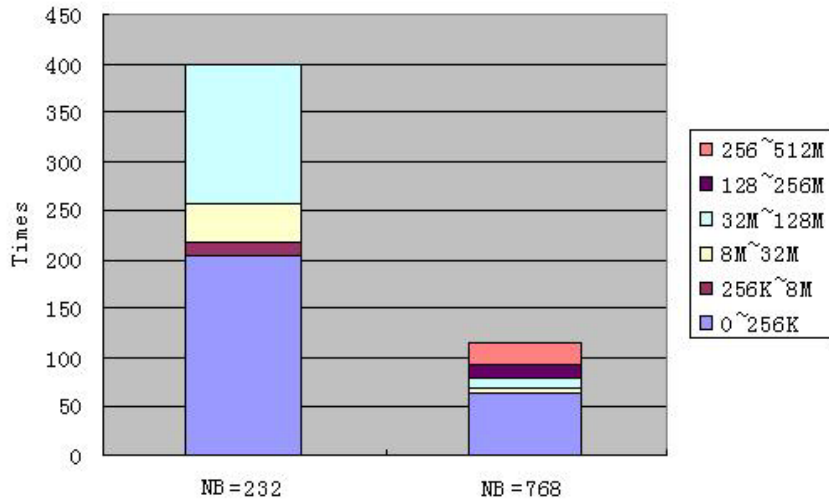
# The Partition in the spotlight



Nebulae@Tianjin, May 2010



# Analysis of communications



Pros and cons of GPGPU for communication system

Pick larger NB

Optimize the perf of large packet communication

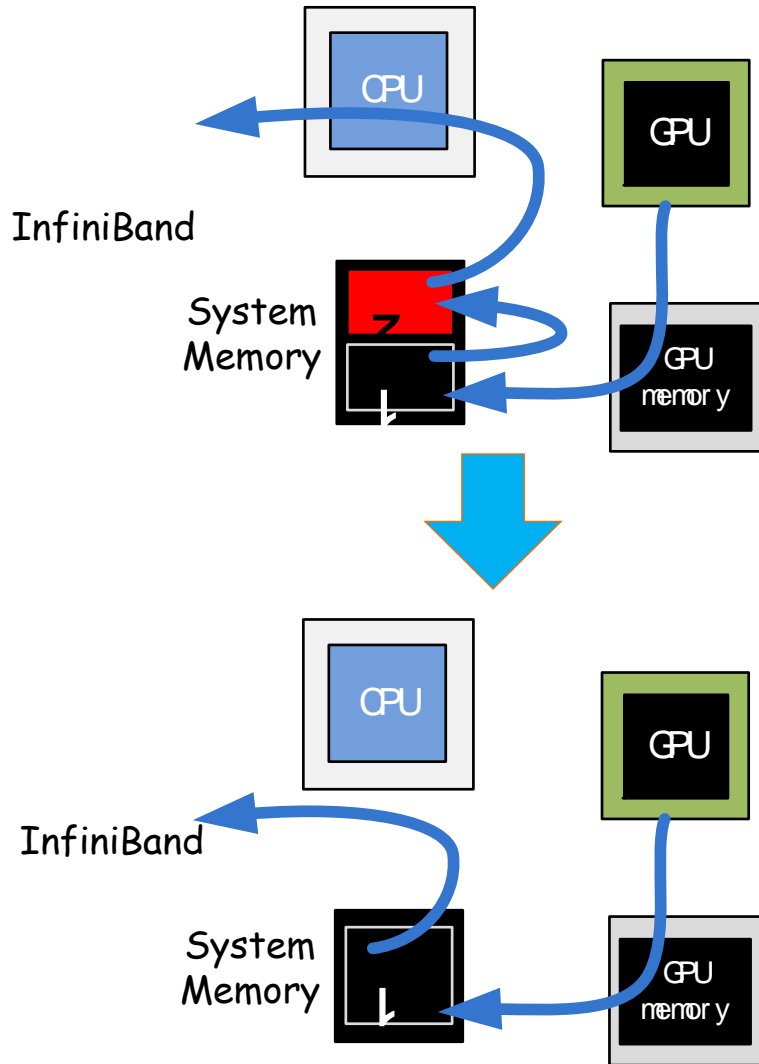
RDMA

Huge-page support

Intra-node optimizations



# GPU-Direct



Why GPU-Direct?

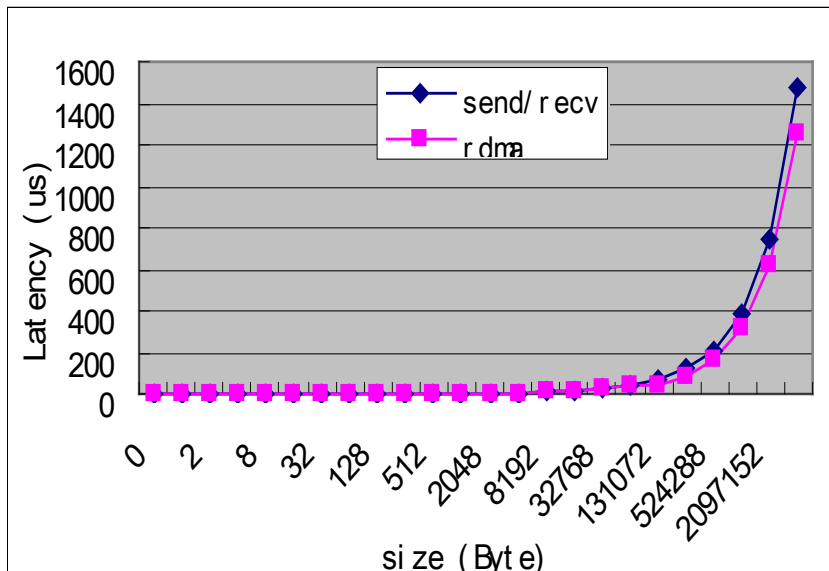
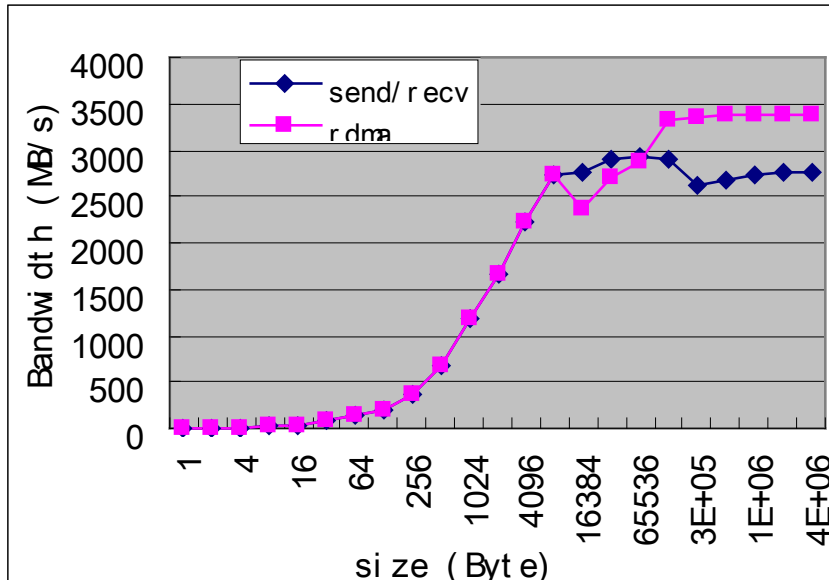
The solution

Acknowledgments

Nvidia

Mellanox

# GPU-Direct(cont.)



Micro-benchmarking results

What next?

Reduce memory reg. and pin-down cache miss rate

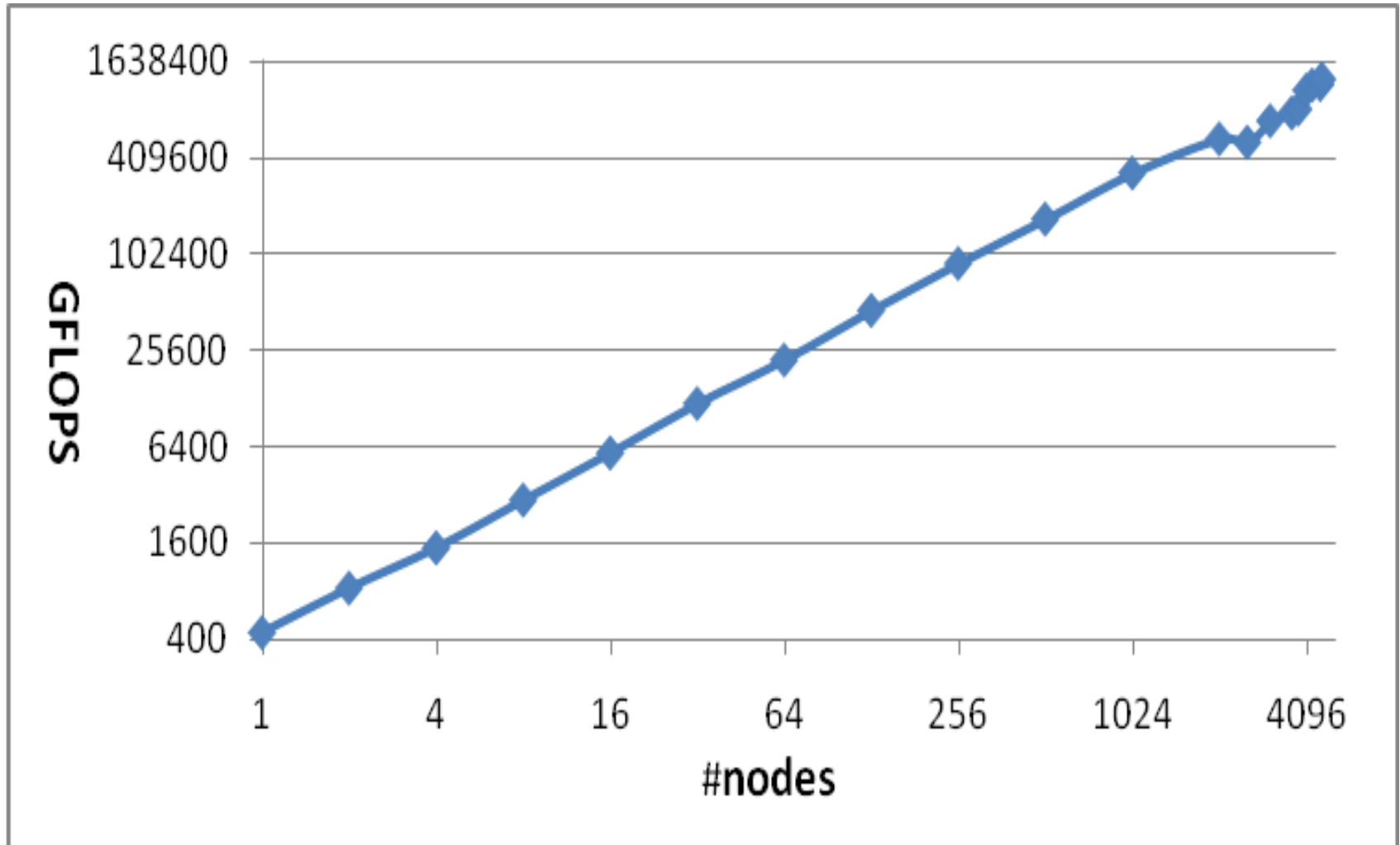
Handle noncontinuous packets more efficiently

Linpack + 1%

using optimized openMPI



# Scale smoothly...



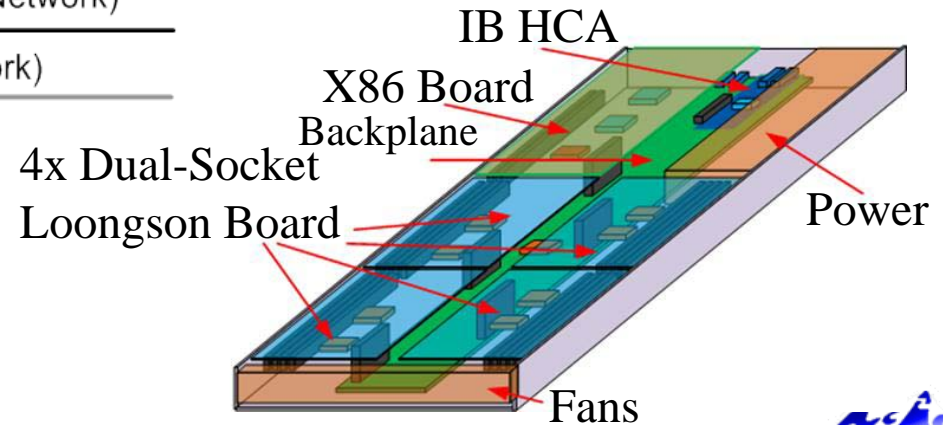
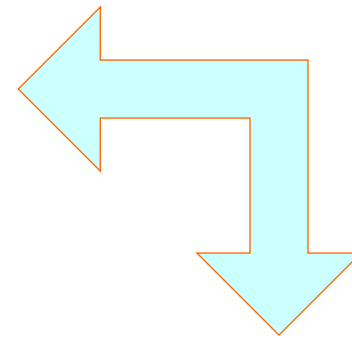
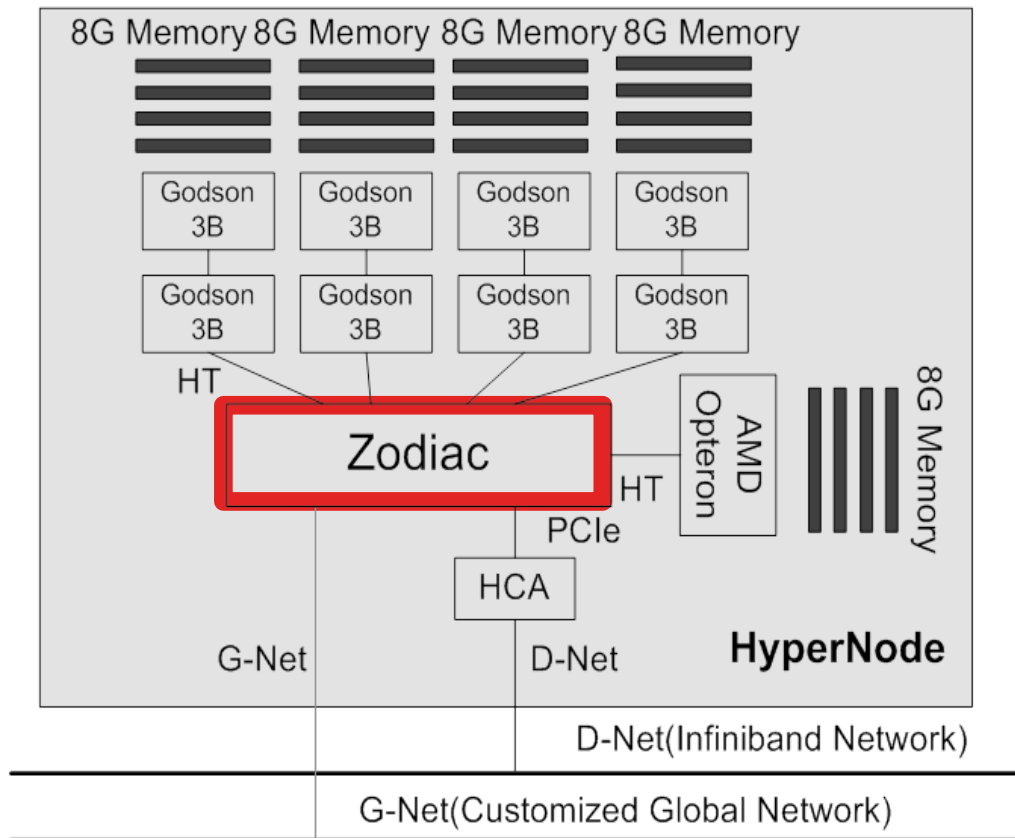
Linpack scores on Nebulae system

---

Communication System  
in  
the  $L^3$  Partition



# HyperNode Architecture



# Design Philosophy

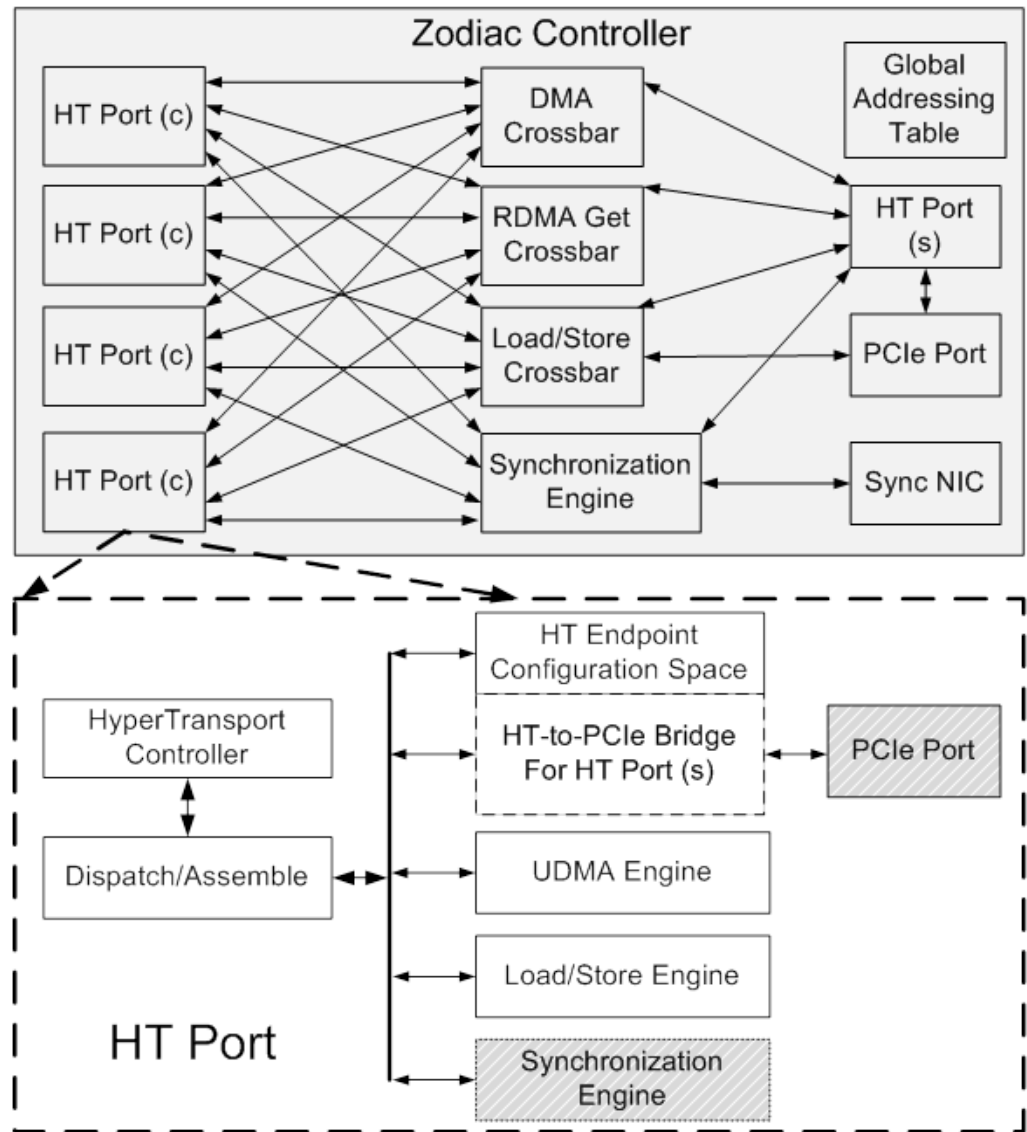
---

- Keep the hardware **simple**.
- Leave the dirty works to software.
- Make it work, then make it fast, and then make it large scale.

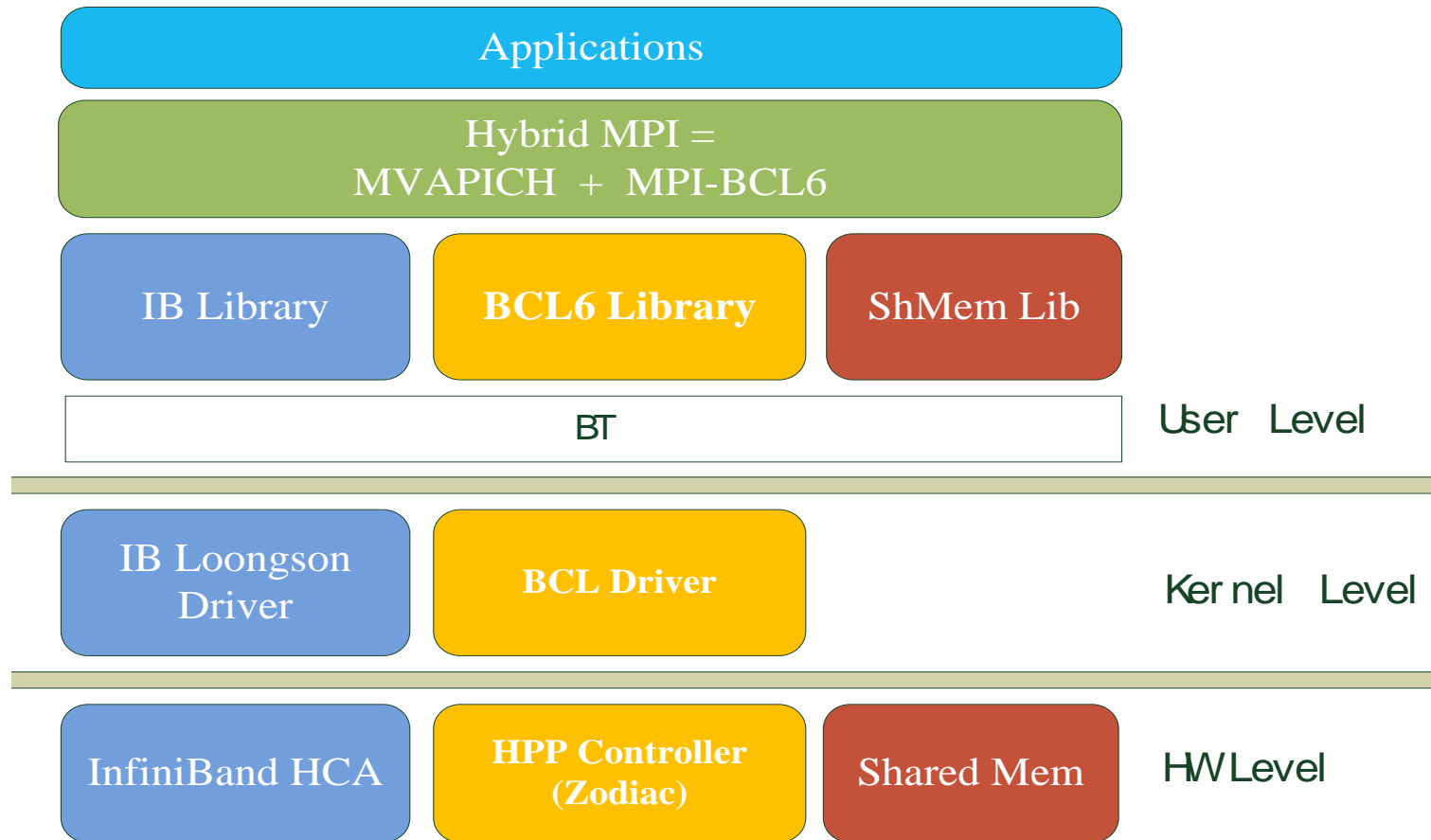


# The system controller: Zodiac

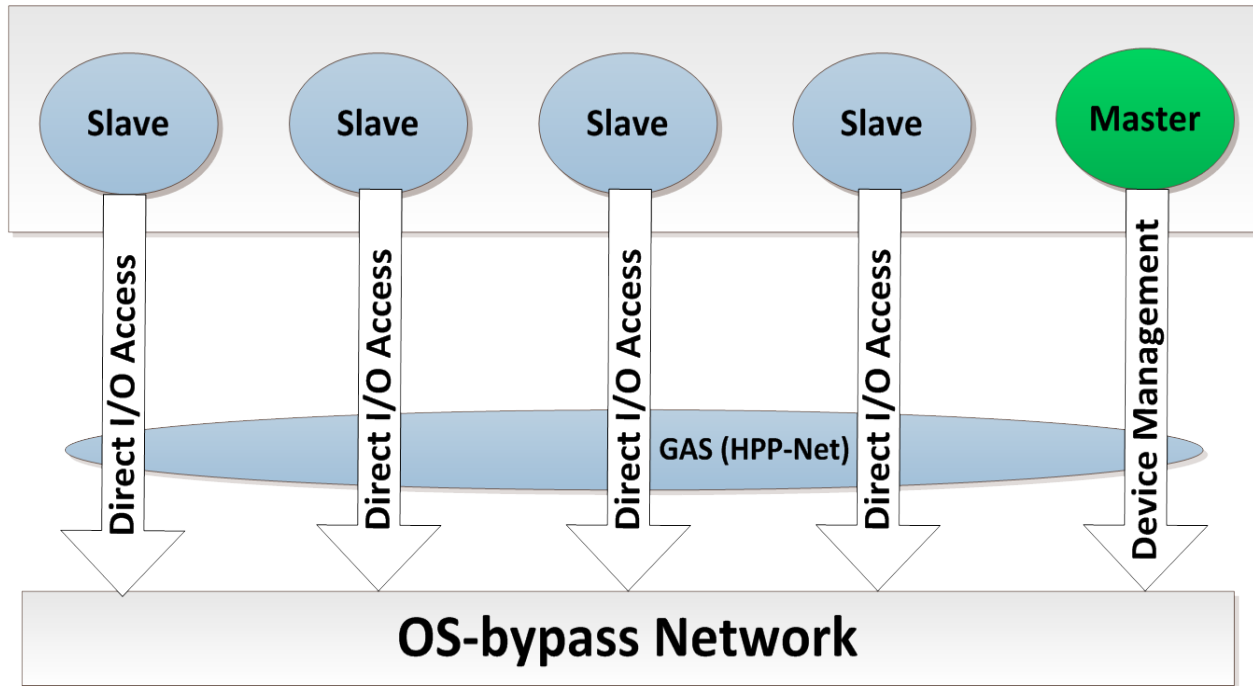
- ✓ Four HT ports
- ✓ No cache coherence
- ✓ Node-wide GAS
- ✓ Multi-mode support
  - ✓ Remote Ld/St
  - ✓ NAP
  - ✓ RDMA
  - ✓ Sync Op.



# The communication software stack

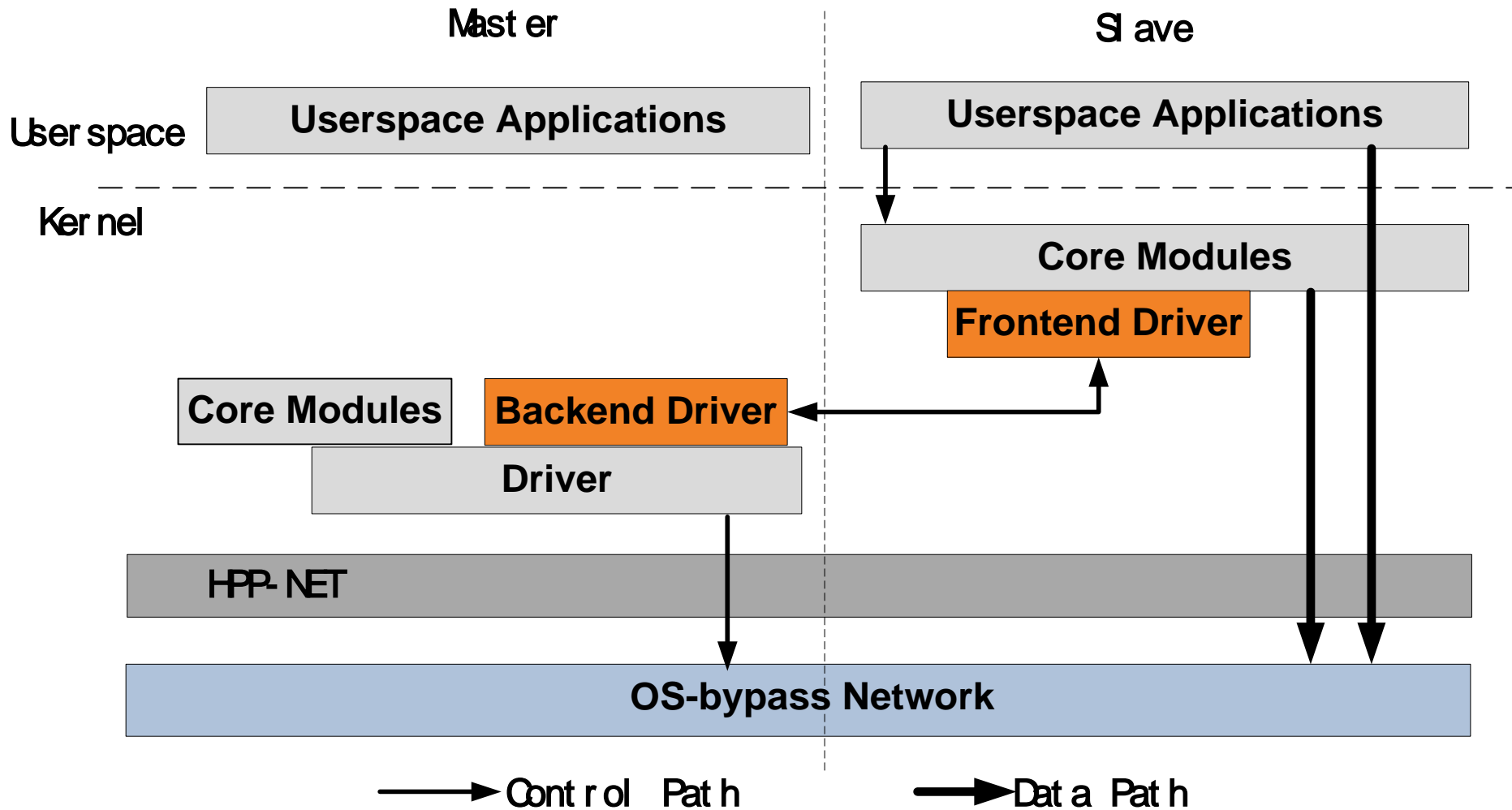


# Inter-node Design



IB Virtualization

# Inter-node Design (Cont.)



IB Virtualization

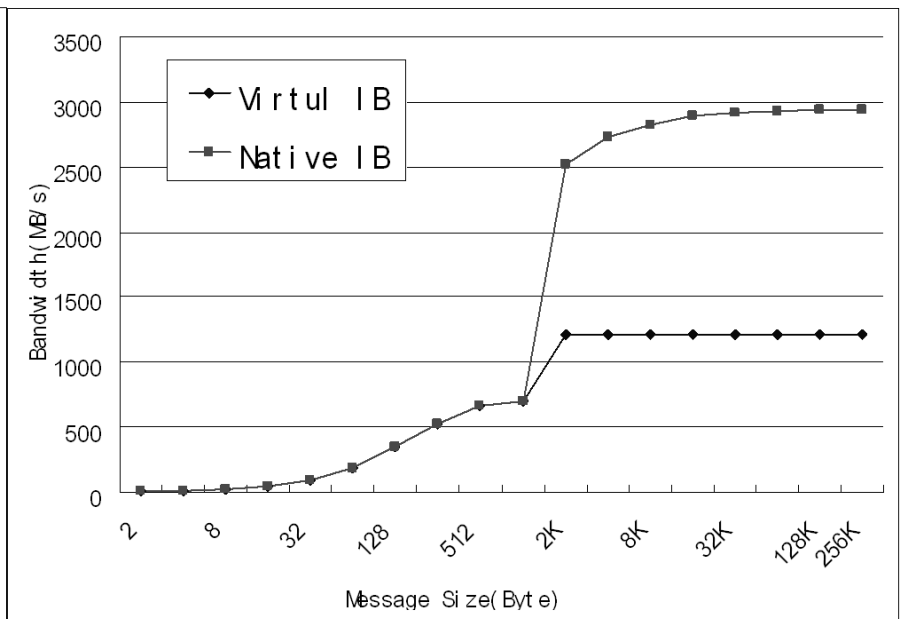
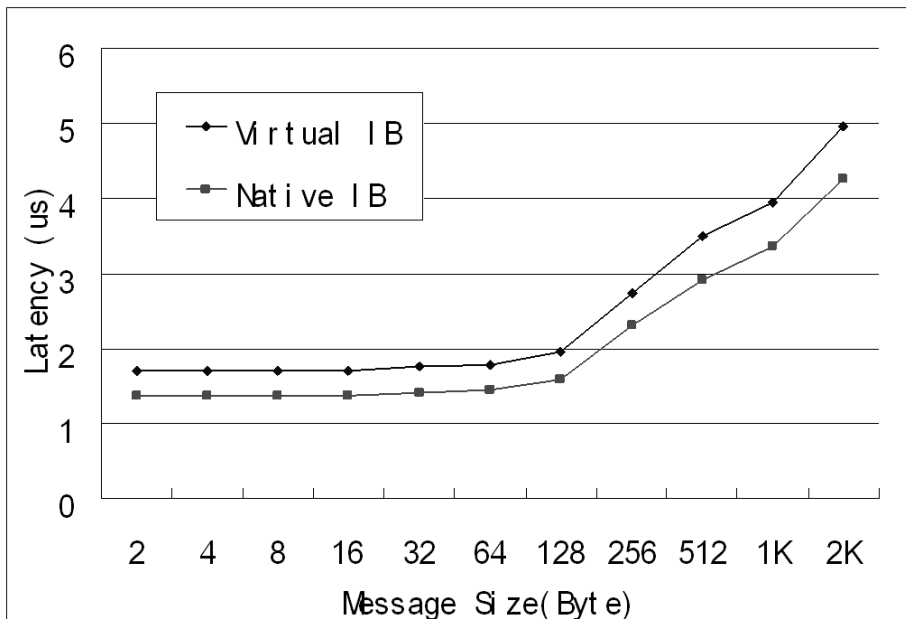


# Inter-node performance

The VIB approach is proved.

BW is bounded by the HT freq.

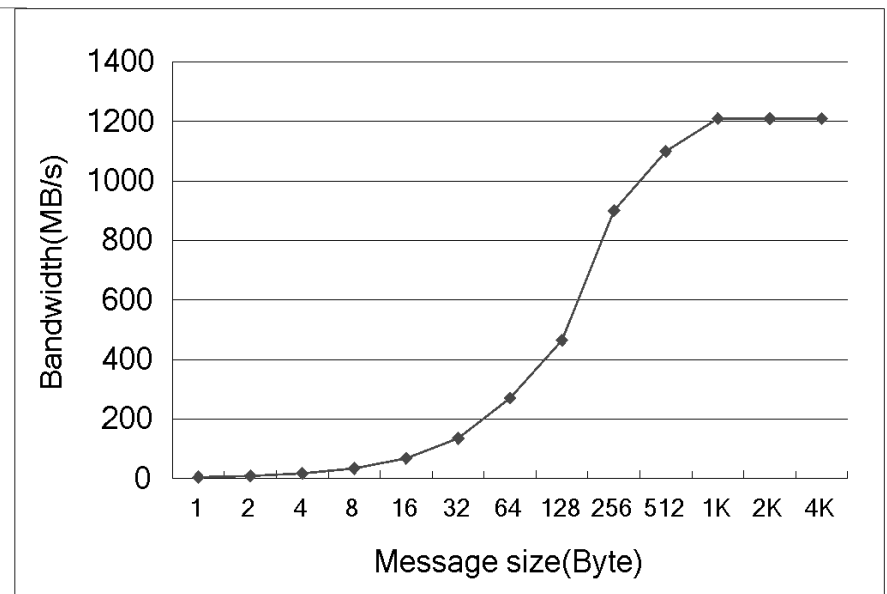
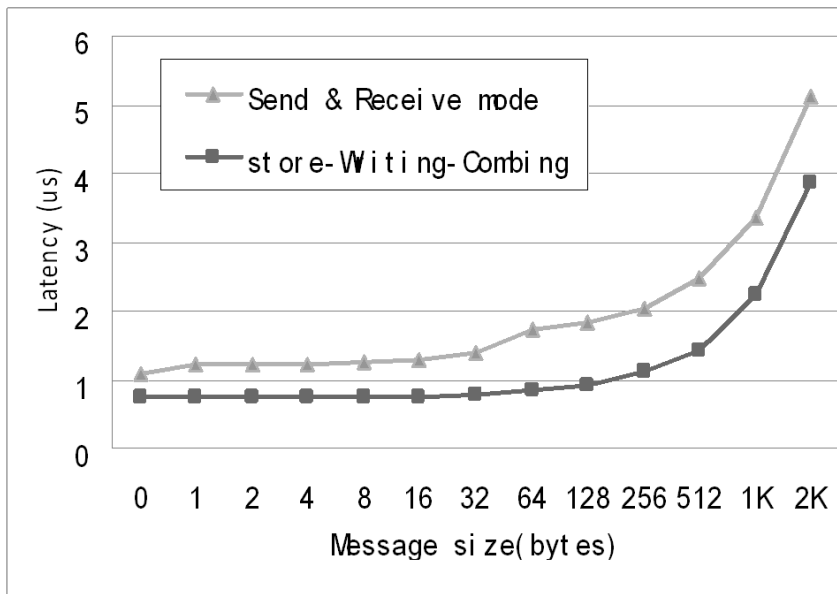
Optimization is still necessary.



# Intra-node performance

The FPGA-based performance

Expecting a large improvement from ASIC version



# Acknowledgments

---

## Hardware Team

An Xuejun, Cao Zheng, Chen Fei, Hu Nongda, Liu Like, Liu Tao, Shen Hua, Wang Dawei, Wang Kai, Wu Dongdong, Xie Liwei, Yang Jia, Yang Xiaojun, You Dingshan, Yuan Guojun, Zhang Peiheng.

## System Software Team

Jiang Tao, Li Bo, Li Qiang, Ma Can, Miao Yanchao, Tang Hongwei, Yang Chao, Yu Huang, Zhang Panyong.

## Application Optimization

TAN Guangming



---

# Q&A

