National Institute of Information and Communications Technology

SC08

Booth Display & Demonstration @ #3003

MaSTER-1: 5-port 10GbE Testbed
Performance Optimization of TCP/IP
GRAPE-DR

Katsuyuki Hasebe

The University of Tokyo
&
National Institute of Information and Communications Technology

Booth Display & Demonstration

- MaSTER-1
  - 5-port 10GbE Testbed
- Stream Harmonizer
  - Optimizing parallel TCP stream/performance
- Performance optimization of TCP/IP
- High speed TCP communication experiments
- CosmoGrid
- GRAPE-DR processor chip/system

MaSTER-1

5-port 10GbE Testbed

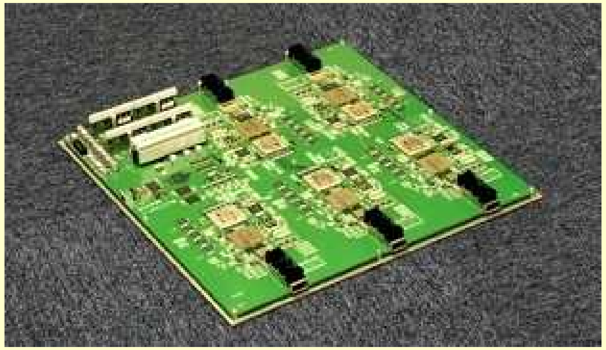
MaSTER-1 Overview

•10GbE (LAN PHY) experimental testbed
•5 XFP ports connected to FPGAs through MACs
•All FPGAs are connected all-to-all at the speed of 10 Gbps
•FPGAs can be communicated with a control PC thorough a USB port

MaSTER-1 Advantages

•Packets input from a port can be processed at the speed of 10 Gbps
•Packets can be output from any ports
•Each port has powerful configurable FPGA
•Each port has large memories to store packets

MaSTER-1 (12 Layers PCB)



MaSTER-1 Applications

• Programmable 10GbE switch
    •Currently running
•Packet Filters
•Packet Logger
•Pseudo Long Fat-pipe Networks
    •Maximum 400 ms delay

MaSTER-1 improves the performance of Parallel TCP Streams

•Multiple ports allow MaSTER-1 to handle multiple connections without switches
•MaSTER-1 can observe directly the packets transmitted by end hosts
•MaSTER-1 will clarify the problems with the method for dropping and merging packets in 10GbE switches
• MaSTER-1 is a good tool for verifying the performance of parallel TCP streams on Long Fat-pipe Networks (LFNs)

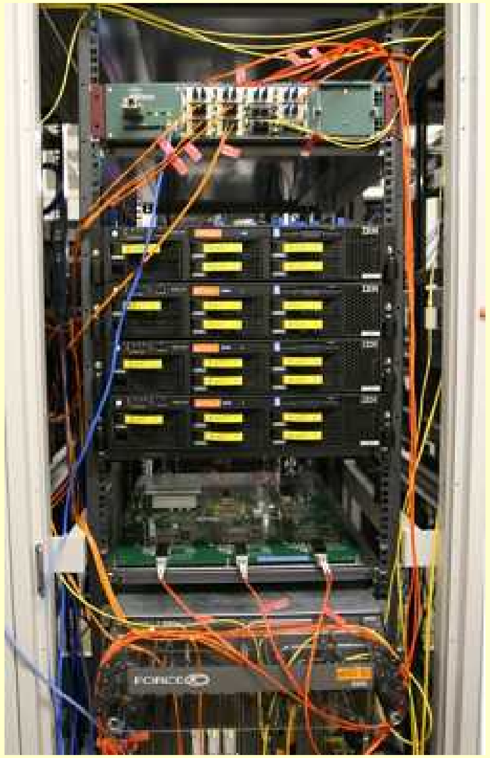
MaSTER-1

5-port 10GbE Testbed

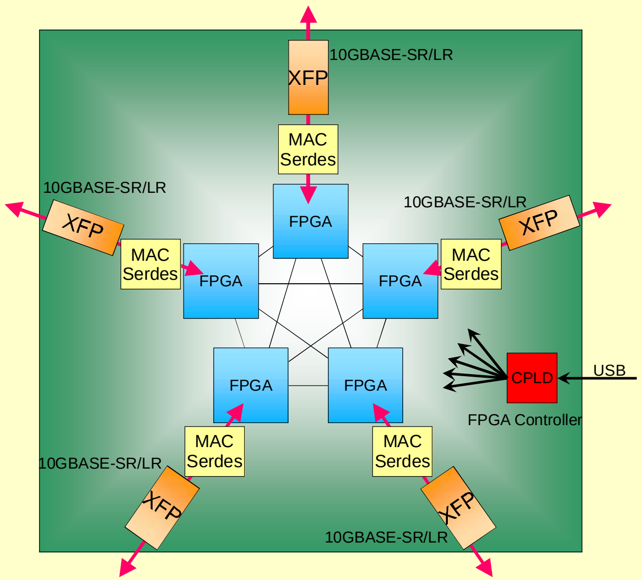
MaSTER-1 Specification

Ports	5 XFPs (10GBASE-SR/LR)
Processors	5 FPGAs (Xilinx XC5VFX70T-1FF1136) – 1 for each port
Interconnection	All-to-all 10 Gbps – Xilinx Rocket IO
Memory	DDR2 SDRAM 2.56 GiB – 512 MiB for each processor
I/O	USB 2.0
Dimension	430 mm x 430 mm x 50mm (WDH)

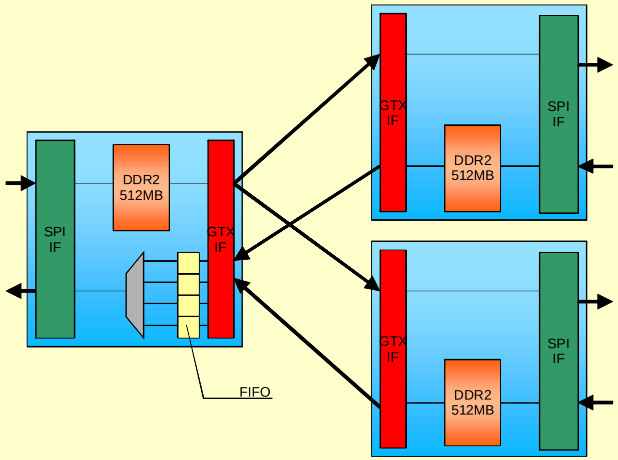
MaSTER-1 in a Rack



MaSTER-1 Block Diagram

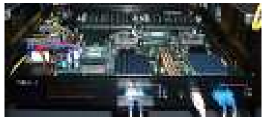


MaSTER-1 FPGA Interconnections



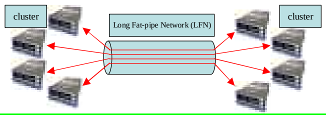
Stream Harmonizer

optimizing parallel TCP stream/performance

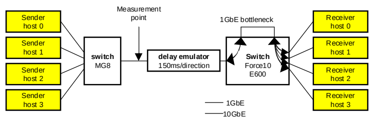


- Inter-cluster data transfer using parallel TCP streams
- Throughput unbalance among streams
  - Fairness lost because of slow recovery of flow on LFN

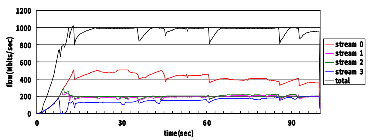
⇨ Data transfer speed limited by slow streams



Throughput unbalance among TCP streams

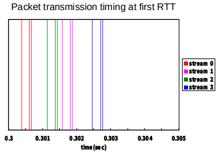


Experimental network to model LFN
- 10 gigabit Ethernet (10GbE)
- Delay emulator
  - 150*2 = 300ms round trip time (RTT)

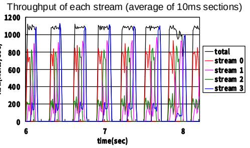


Throughput of each stream (average of 300ms sections)
- Different packet loss timing at starting phase
  - Generating throughput unbalance among streams
- Slow feedback to shrink the throughput difference
  - Throughput unbalance preserved

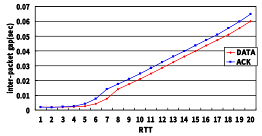
Packet transmission timing at first RTT



Throughput of each stream (average of 10ms sections)

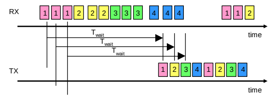


- Packet transmission order preserved among streams
  - stream 0, 2, 1, 3
- Last stream experiences packet loss first
  - switch buffer crowded by the other streams
  - packet loss order : stream 3, 1, 2, 0



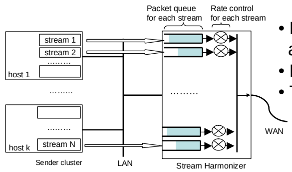
Packet transmission timing difference between stream 0 and 3 in each RTT
- Difference increases gradually
  - Increased packet gap by the 1GbE bottleneck
  - ACK timing preserved in the next RTT
- Initial transmission order preserved with increased inter-packet gap

Packet scheduler to synchronize loss timing



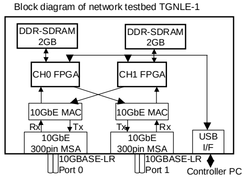
- Throughput unbalance because of different loss timing

⇨ Synchronizing packet loss timing by shuffling them

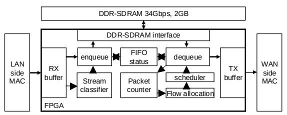


- Packet scheduling hardware at sender gateway
- Per-stream queue
- Transmission timing control for each stream

Block diagram of network testbed TGNLE-1

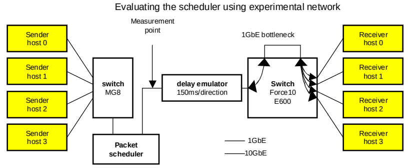


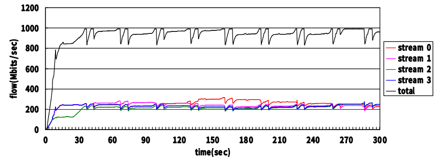
Packet scheduling FPGA logic on TGNLE-1



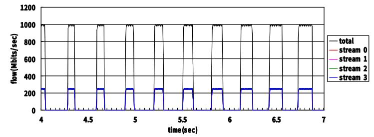
Flow balancing effect of the proposed scheduler

Evaluating the scheduler using experimental network





Throughput of each stream (average of 300ms sections)
- Synchronized packet loss timing
- Less throughput unbalance among streams



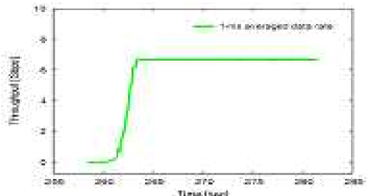
Throughput of each stream (average of 5ms sections)
- Synchronized packet transmission
- Fairness among streams

NiCT

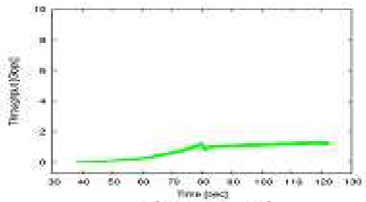
Performance Optimization of TCP/IP

Rapid expansion of Long Fat pipe Network (LFN) all over the world.
10Gbps network interfaces are available for commodity PCs with reasonable price.
Our objective is to fully utilize 10Gbps connection with TCP/IP communication.
External bus speed such as 10GbE becomes comparable with internal.
There exist many causes of performance decrease of LFN TCP communication.
Complex of these causes makes situation more complicated.

Mysterious phenomena



PCI-X NIC

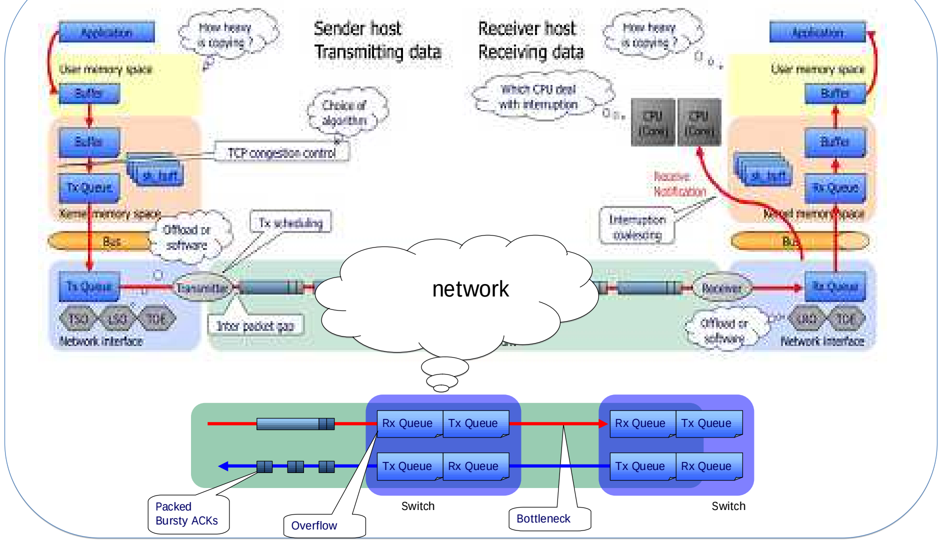


PCI-express NIC

Early version of 10GbE NIC uses PCI-X internal bus, less than 8Gbps.
Version up of NIC with PCI-express (max speed is full10Gbs) causes terrible performance decrease.

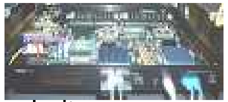
Utilizing hardware support for TCP communication also brings strange phenomena, such as slow-down of scaling speed.

Possible Causes of Performane Decrease



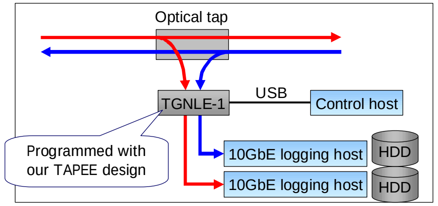
Performance Optimization of TCP/IP

TAPEE

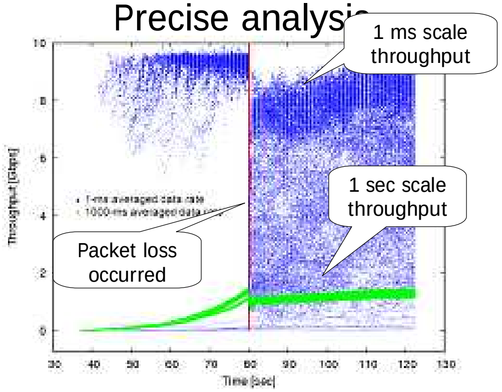


TGNLE-1
[Sugawara '05]

- Analysis with fine time granularity
  - 100 ns
- Raw behavior of packets on a network

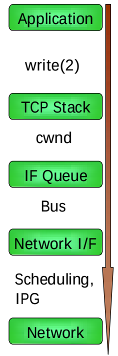


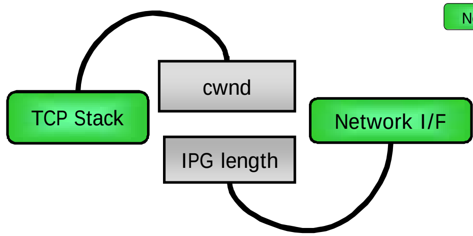
Clarification of Issue by Precise analysis



Packet Pacing (1)

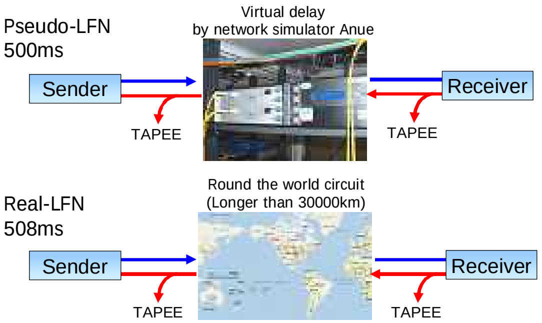
- Compared and evaluated methods
  - X Application level
  - X Limiting window size
  - O Inter packet gap (IPG) control [Chelsio]
- IPG control is essential for WAN PHY
  - 700 octet NG, 720 octet OK
- Without pacing, TCP cannot achieve congestion avoidance



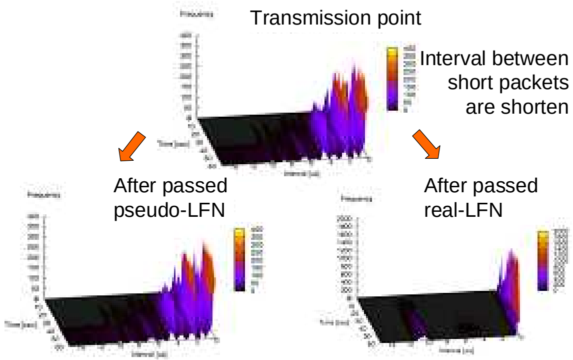


Dynamic Adjusted IPG Control

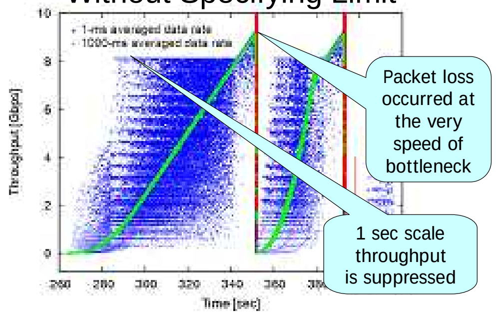
Comparison of Pseudo and Real LFN using TAPEE



Pseudo vs. Real



Successfully Probed the Bottleneck Without Specifying Limit



NICT

Internet2 Land Speed Record

- 99% of physical bandwidth for 5 hours on 522ms RTT network

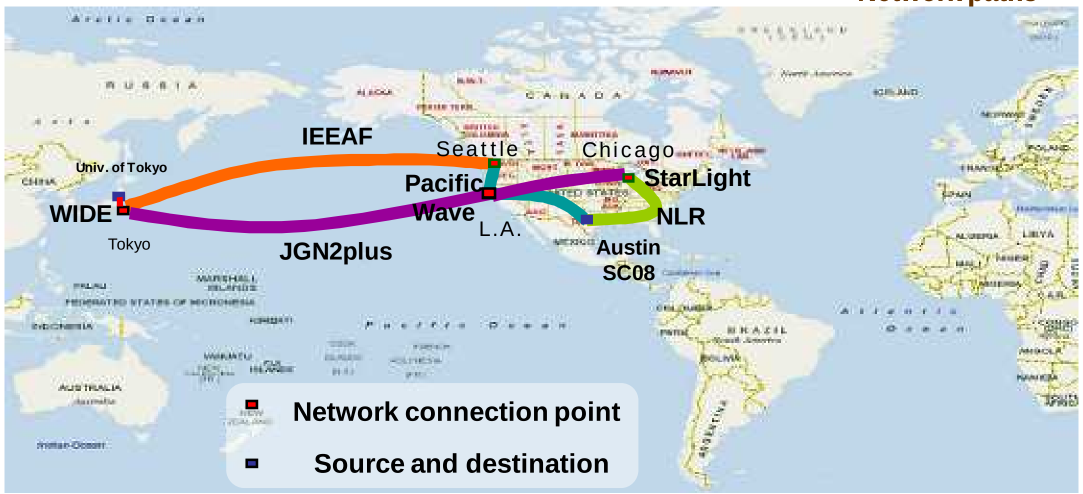
Thank you

High-speed TCP communication experiments

- Goal
  Efficient TCP communication on Long Fat pipe Networks
  - Single and multiple stream TCP
  - Adaptive inter-layer cooperation
  - Balancing parallel TCP streams
  - Austin(SC08) ⇔ the University of Tokyo
- Network
  - WAN PHY 10 Gbps network
  - 9.2 Gbps maximum payload performance
- System used
  - Intel IA32 servers with Chelsio S310E network adaptor
  - MaSTER TCP stream stabilizer
  - TAPEE network instrumentation device

Destination: Faculty of Science, the University of Tokyo

Network paths



CosmoGrid

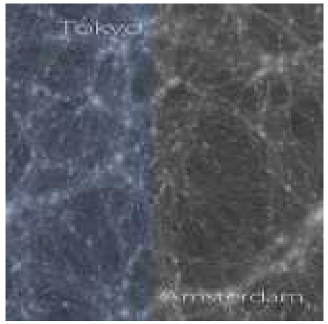
- Using 30 processors on Cray XT4 of National Astronomical Observatory of Japan
- 30 processors of Huygens(IBM Power6) cluster of the University of Amsterdam.
- Cosmological N-body calculation with 256^3 particles
- The size of the simulation box is 60Mpc (megaparsec), with comoving coordinates and periodic boundary
- The specialized calculation code was developed to reduce the required communication bandwidth between two computers and to allow for large communication latency.

Cray XT4 at Tokyo

Huygens (Power6) at SARA

Simulation Results (Test run)

Blue part on Cray, Gray part on IBM



GRAPE-DR Processor Chip/System

GRAPE-DR Processor chip

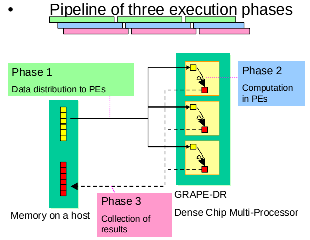
The University of Tokyo
National Astronomical Observatory of Japan

Specification

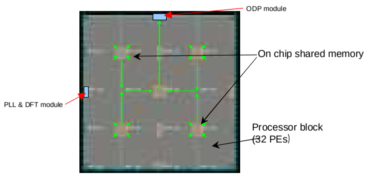
Technology: 90nm CMOS
Number of PEs 512 PEs
Peak performance
512Gflops(single)
256Gflops(double)
Size: 18mm X 18mm
Number of Tr about 400M Tr
Clock Freq. 500Mz
Power Consumption
MAX 60W
Idle time 30W

Pipelined program execution

- Pipeline of three execution phases

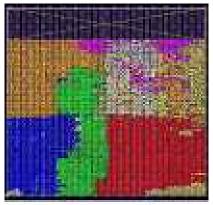


Floor plan of the chip



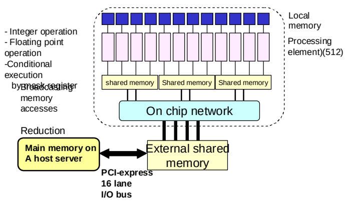
Layout of a PE

Module Name	Color
Register	red
Flt. Mul	orange
Flt. Add	green
Int ALU	blue
Local Memo.	magenta
others	yellow



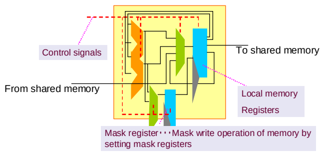
Processor chip architecture

- SIMD like architecture
- 512x64 bit arithmetic units + shared memory + broadcast/reduction network
  - Elimination of inter-PE interconnection
  - Dedicated Reduction network (with arithmetic operations)



Processing Element

- 512 PE in a chip



This research is partially supported by the Special Coordination Fund for Promoting Science and Technology from Ministry of Education, Culture, Sports, Science and Technology, Japan.

GRAPE-DR supercomputing system(2008)

Specification (2009)

Peak Performance: 2Pflops
Number of SING chips 4096 chips
Number of servers 512
Power Consumption 400KW
Size 40 Racks

Interconnect >10 Gbps
OS Linux 2.6.x

GRAPE-DR compliers

- Optimizing compiler for GRAPE-DR
  - Automatic parallelization by global analysis
  - Special purpose optimization for GDR architecture

○ Compiler Ver.1 flat-C compiler (2005)
・Parallel constructs, parallel statements
・Explicit description of parallelism
○ Compiler Ver.2 Optimizing C compiler
○ Currently, a prototype compiler is working
・ Generate native GRAPE-DR codes

Application fields

◎Highly efficient application fields for GRAPE-DR
- N-body simulation in Astronomy,
- Molecular dynamics (MD),
- CFD (SPH method, Global model) etc.
- Linpack, linear systems

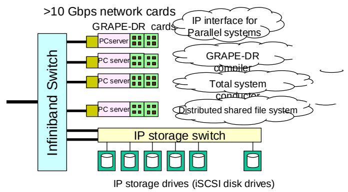
○Application fields with Effective acceleration by GRAPE-DR
- Simulation in nano-technology
- Simulation in bio-technology (FMO etc)

×Application fields with wide memory accesses
- Classical CFD, FFT
- Application software optimized for vector processors

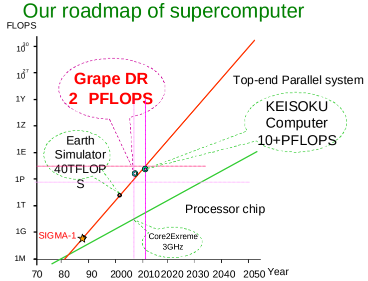
×Application fields with network bottlenecks
- QCD

GRAPE-DR covers about half of important scientific applications

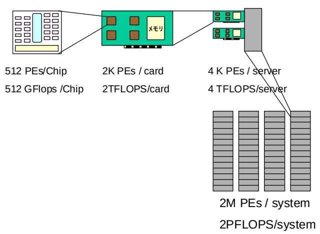
GRAPE-DR system image



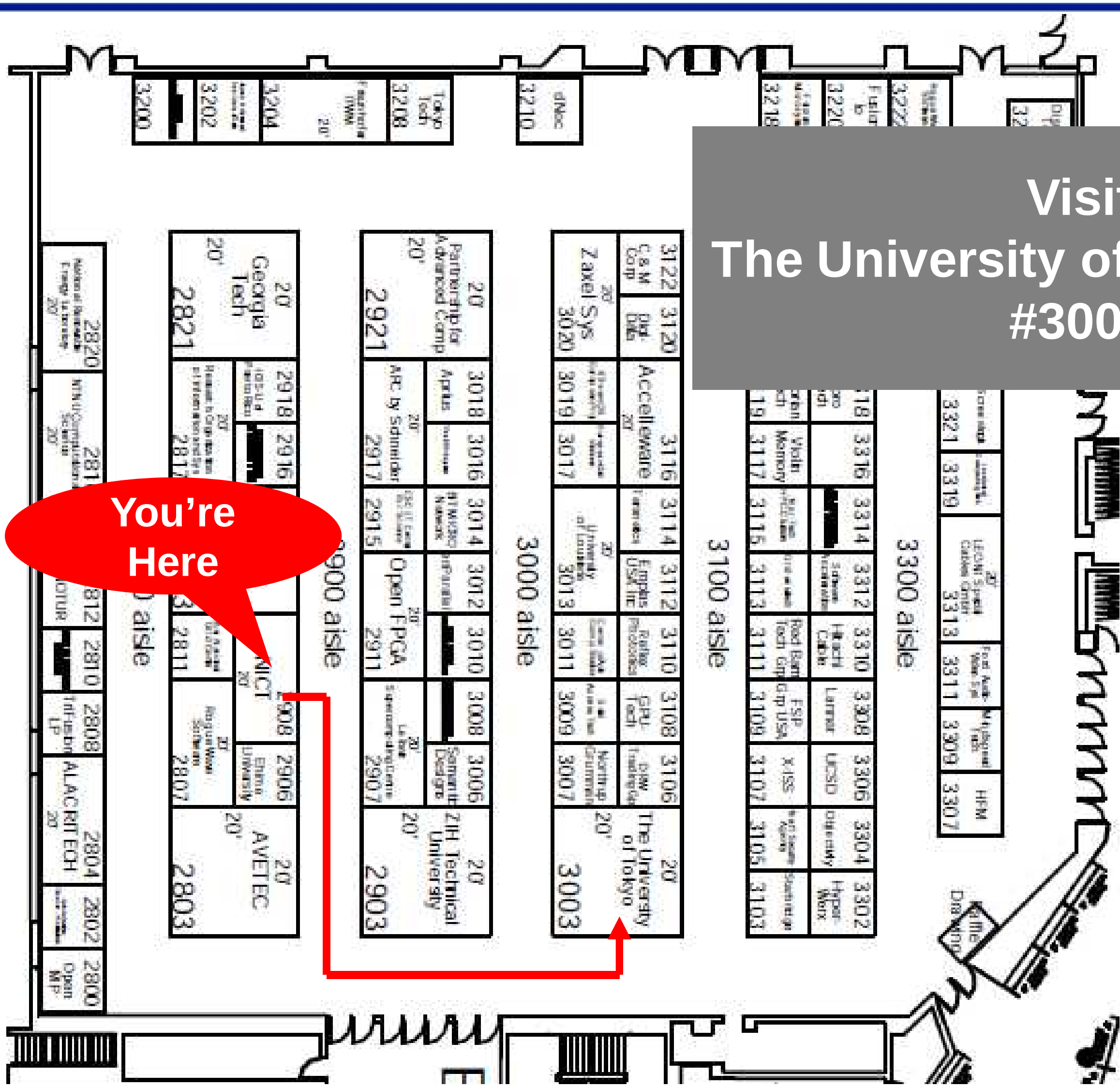
Our roadmap of supercomputer



Hierarchy of GRAP-DR



For more information



Visit
The University of Tokyo booth
#3003