Megascale Project
A Low-Power and Compact Cluster for High-Performance Computing

Hiroshi Nakamura (U. Tokyo)

Masaaki Kondo (U. Tokyo)

Hiroshi Nakashima (Toyohashi UT)

Mitsuhisa Sato (U. Tsukuba)

Taisuke Boku (U. Tsukuba)

Satoshi Matsuoka (TITECH)

http://www.para.tutics.tut.ac.jp/megascale/

Background: Mega-Scale Project (1/2)

- Many applications need Peta-Flops.
  - Computational Genetics/Biology
  - Simulation of Environment/Crimate/Disaster
  - Computational Chemistry/Phisics/…
- Can we achieve Peta-Flops by extending traditional MPP/clusters? ➔ NO!!
  - Huge space requirement (Gym @ 10^4 PE)
  - Huge power requirement (10MW @ 10^4 PE)

➔ We need a new approach!!

=Peta-Flops with Commodity Technology

Background: Mega-Scale Project (2/2)

- Our Mega-Scale project aims to establish fundamental technologies for 10^6 scale parallel systems focusing on;
  - Feasibility to build them with realistic cost and space ➔ **low-power for smaller footprint/volumn**
  - Dependability to operate them with high reliability and fault-tolerance
  - Programmability to obtain maximum performance with minimum effort

  based on commodity technologies.
- about €3M for 5 years, supported by JST (Japan Science and Technology Agency)

MegaProto : Prototype

- Objective : Proof of our claims
  - commodity technology > HPC dedicated
  - low-power/high-density > high-end/low-dens.
- Platform for our software development

  still under development, but…
  - power-aware compilation
  - high-performance/dependable NW: RI2N (Redundant Interconnection with Inexpensive Network)
    - network trunking for performance
    - network redundancy for reliability
  - fault-tolerant cluster management
    - Skewed Checkpointing for Multiple Failures (SRDS'04)

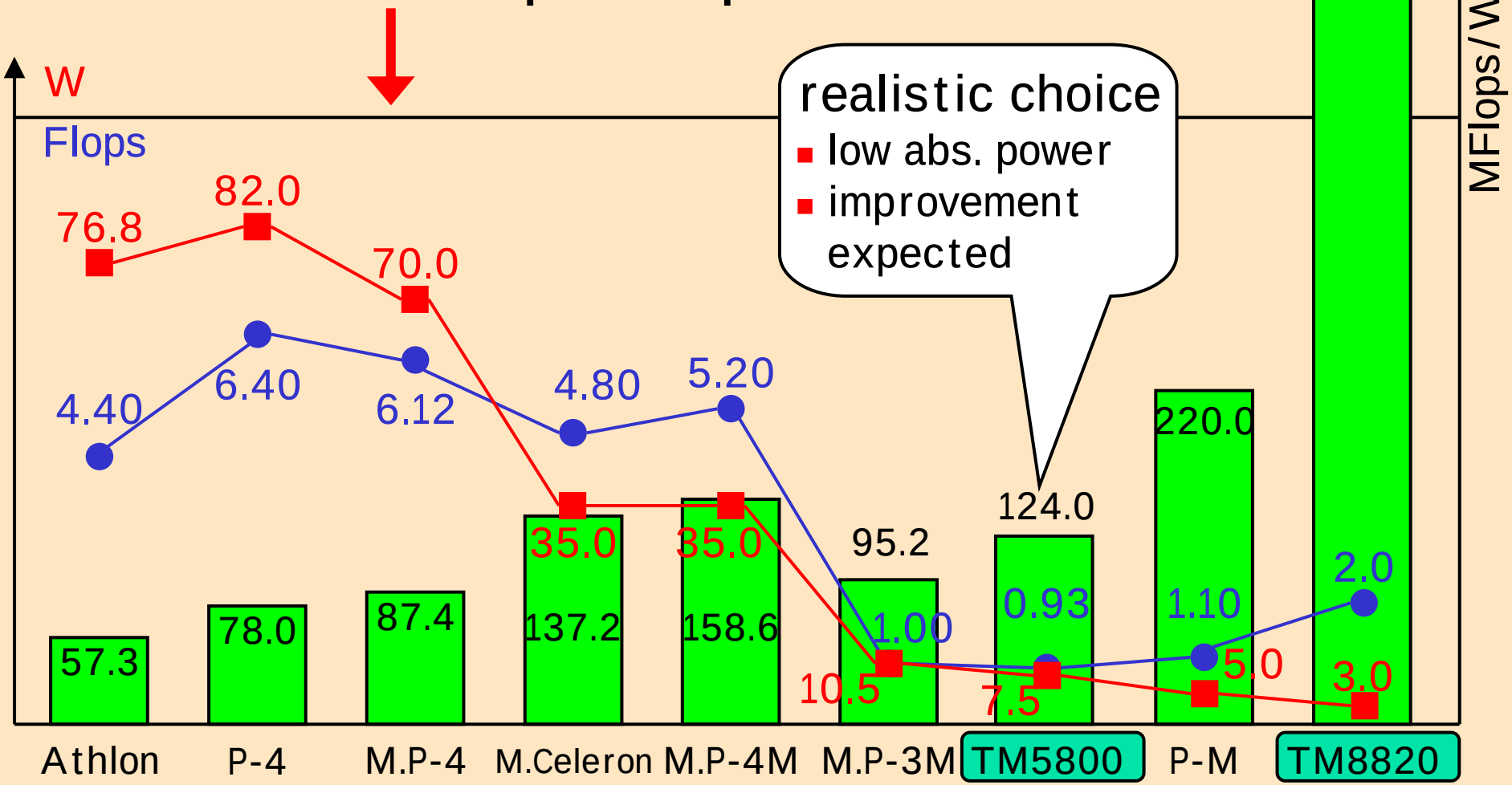
Conceptual Design (1/2)

performance/power perspective

- Target power & perf./ (19" x 42U: 1rack)
  - peak perf. = 1TFlops
  - power = 10kW (300W/1U cooled by air)
  - ➔ perf/power = 100MFlops/W
- Breakdown of power budget
  - processors = 1/4
    ⇒ 400MFlops/W
  - proc peripheral (mem. etc) = 1/4
  - network = 1/2

Conceptual Design (2/2)

- Is 400MFlops/W processor available



System Configuration (1/4)

version 1

- TM5800 (Crusoe)
- 0.93GFlops
- L1C =64KB
  L2C =512KB
- 256MB SDR

65mm

130mm

System Configuration (1/4)

version 1
- TM5800 (Crusoe)
- 0.93GFlops
- L1C =64KB
  L2C =512KB
- 256MB SDR

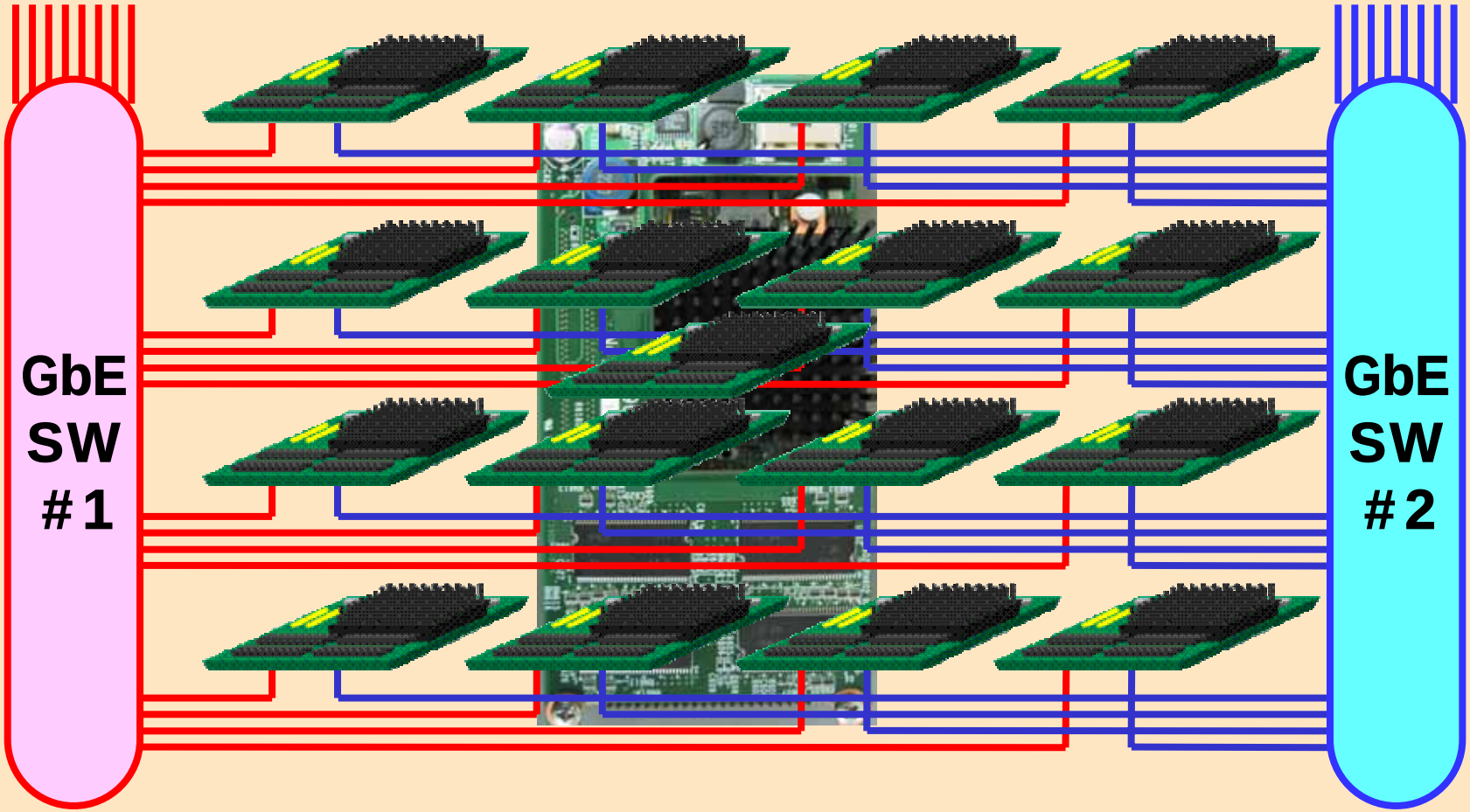
version 2
- TM8820 (Efficeon)
- 2.0GFlops
- L1C =192KB
  L2C =1MB
- 512MB DDR

65mm

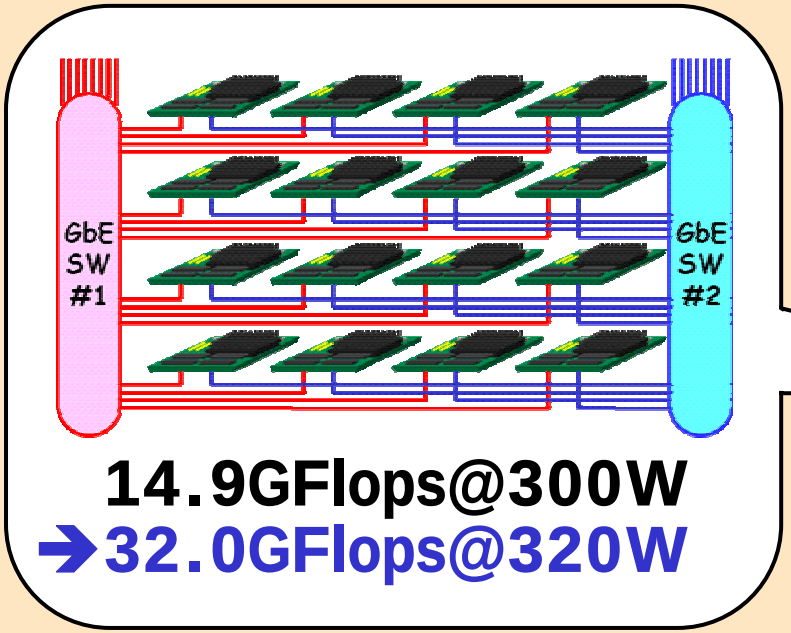
130mm

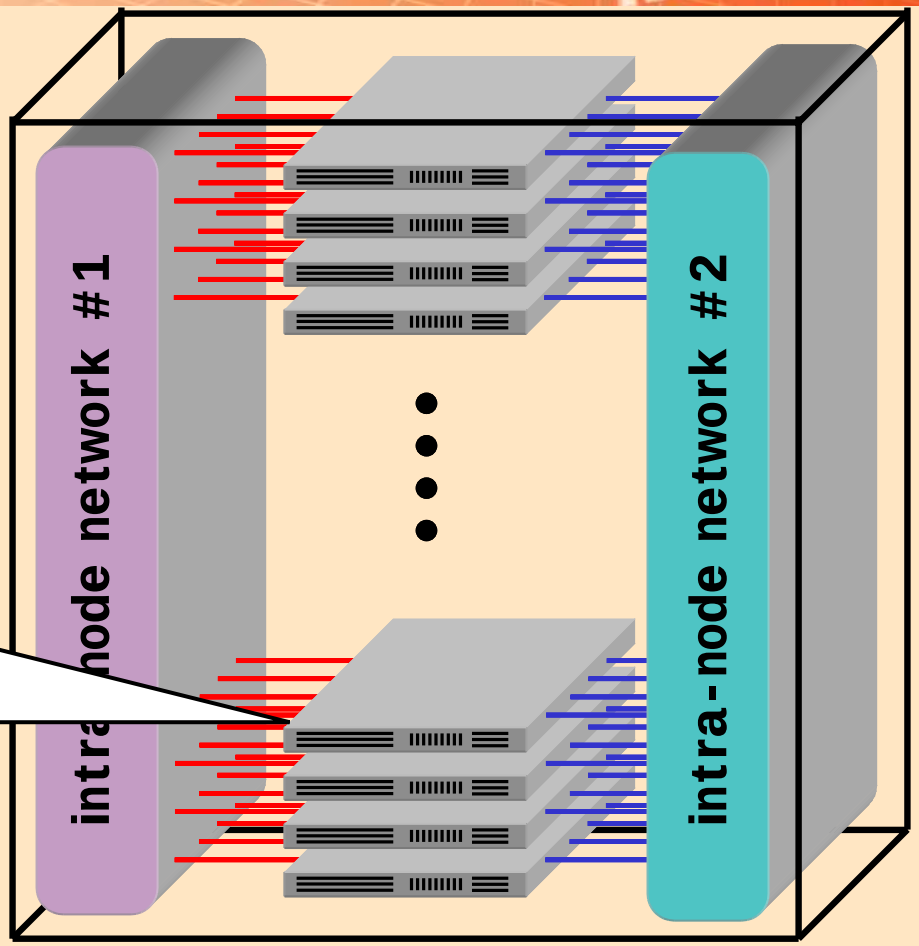
2-stage rocket !!

System Configuration (2/4)



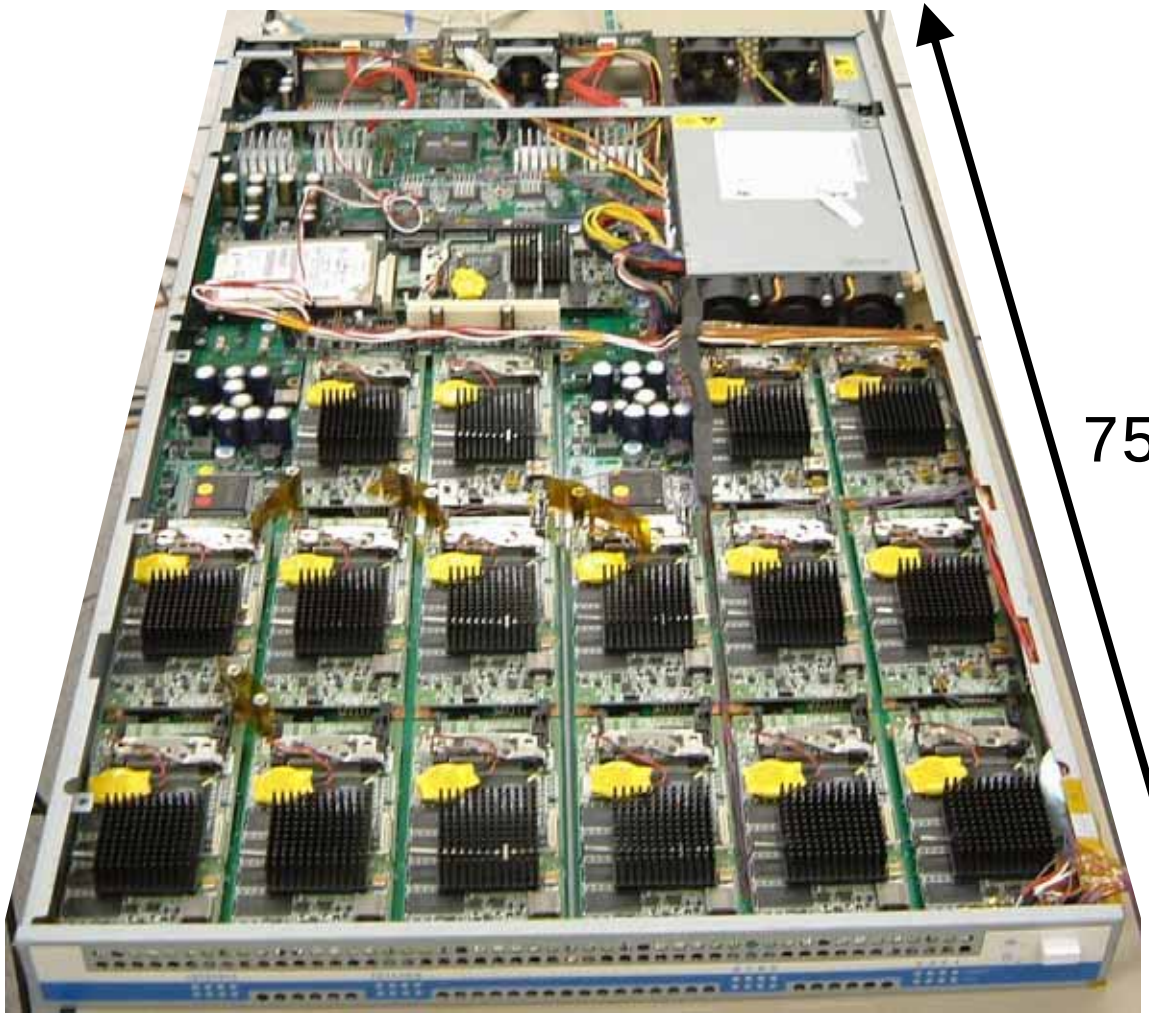
System Configuration (3/4)





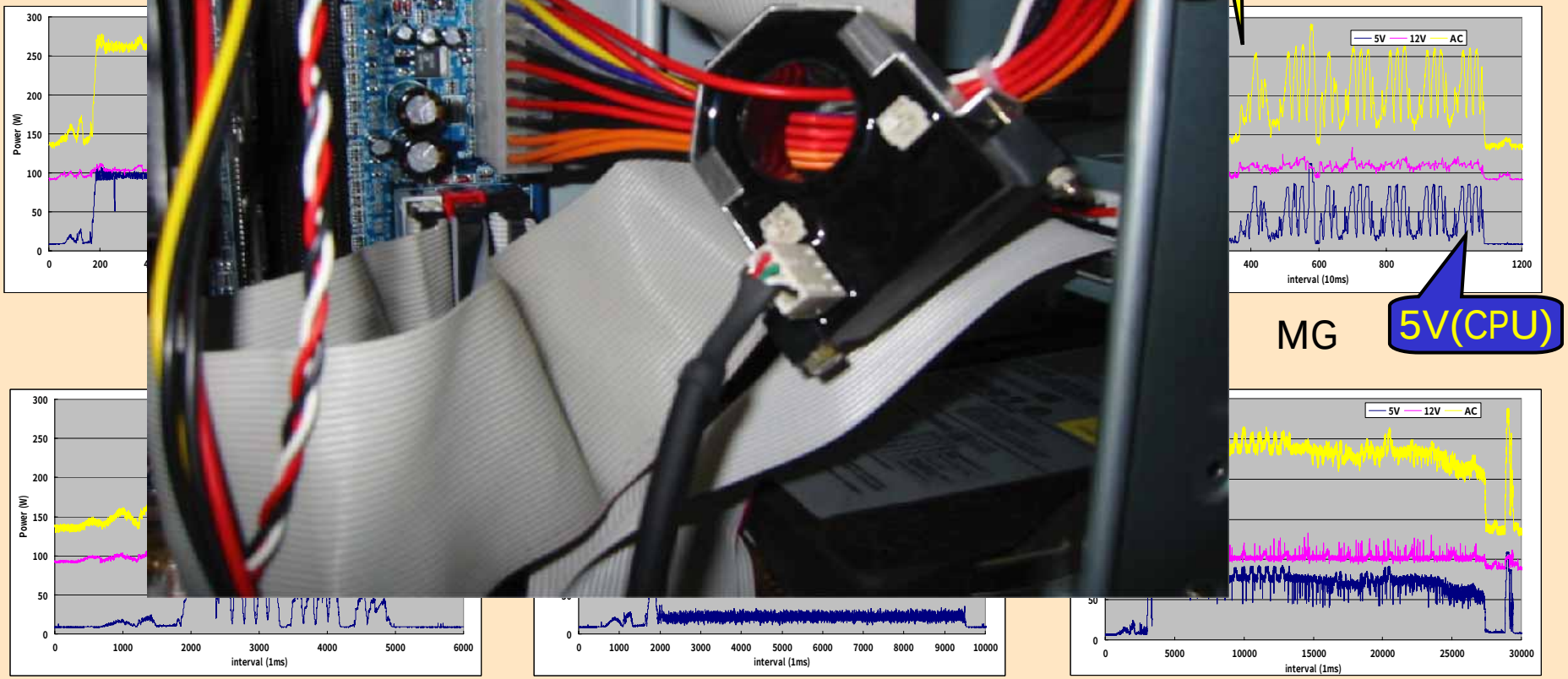
System Configuration (4/4)

version 1, delivered March, 2004

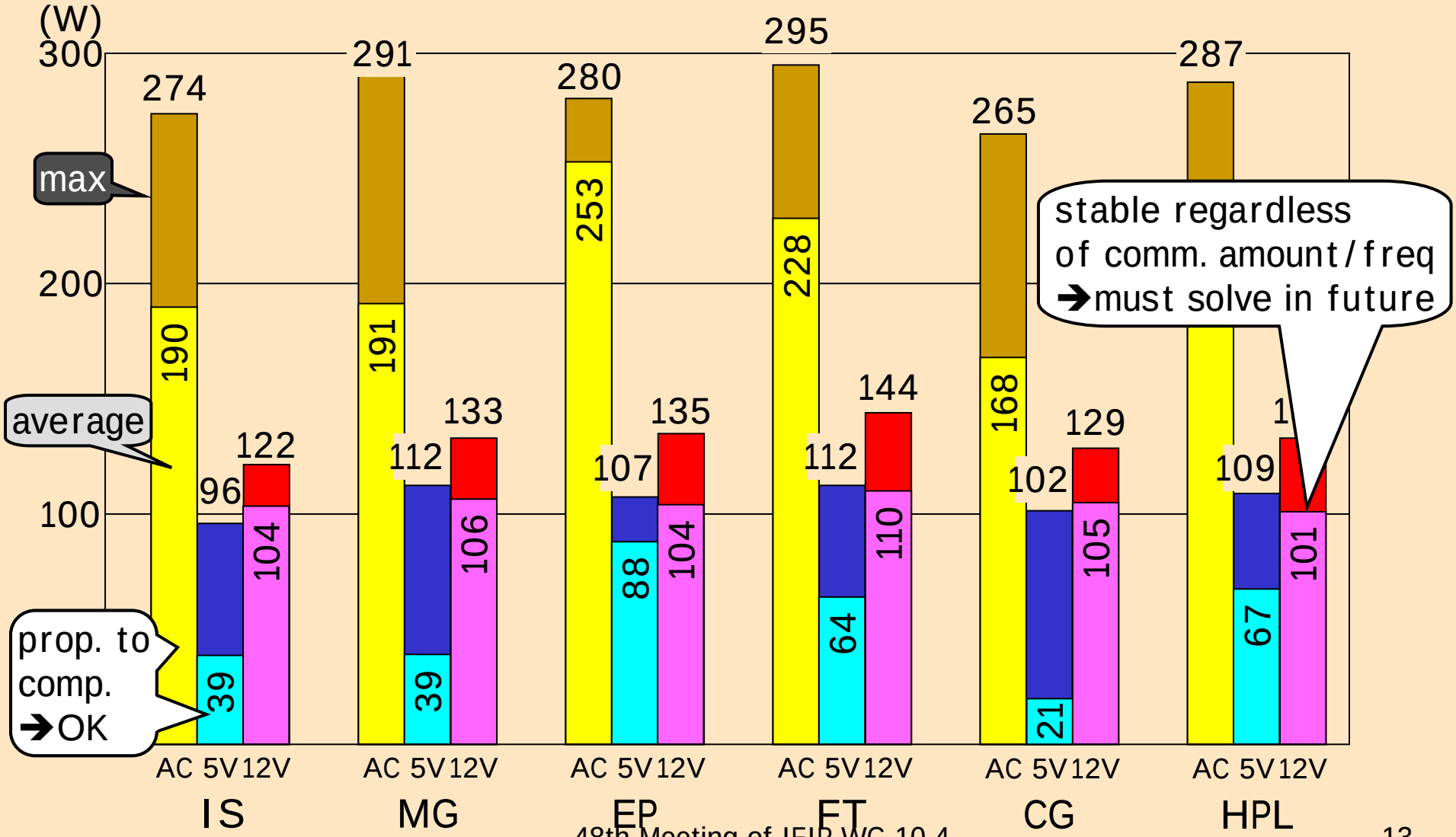


Performance Evaluation of MegaProto/Crusoe (1/3)

- P



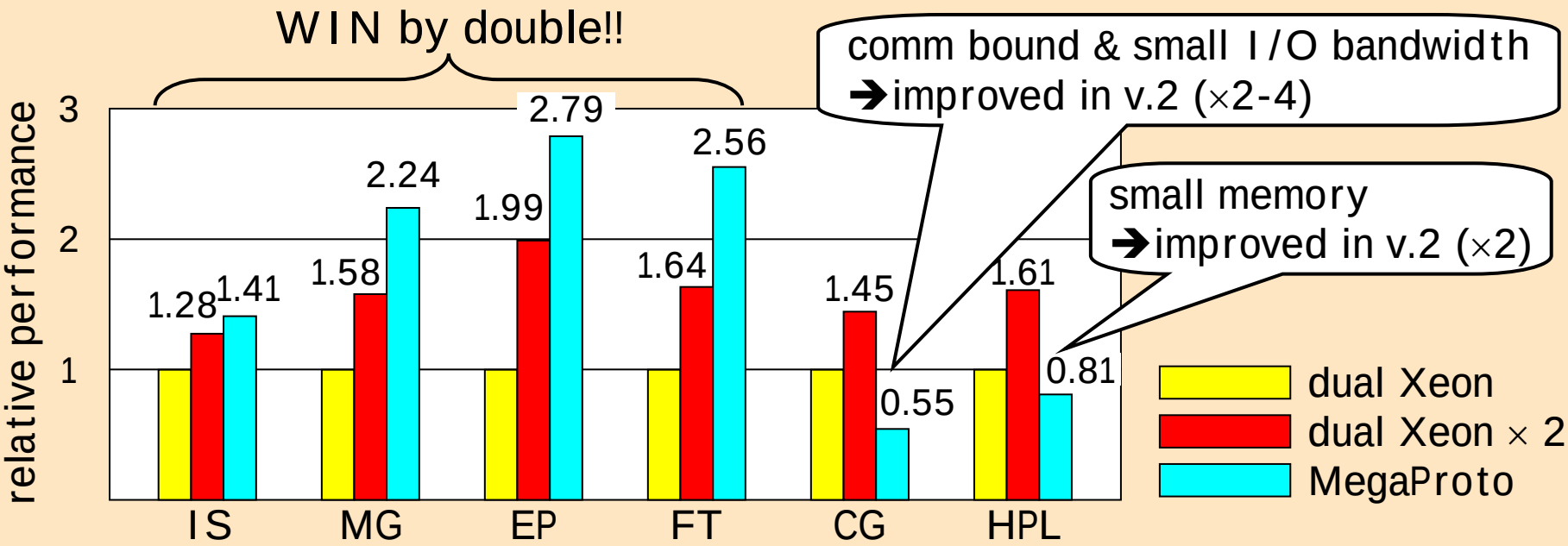
Performance Evaluation of MegaProto/Crusoe (2/3)



Performance Evaluation of MegaProto/Crusoe (3/3)

- v.s. 1U server (dual Xeon 3.06GHz, 1GB)

	dual Xeon	MegaProto
power / 1U	400W	300W
processor TDP	170W	120W
peak perf.	12.24 GFLOPS	14.88 GFLOPS



Summary

- Megascale Project : A Low-Power and Compact Cluster for High-Performance Computing
  - megascale high-performance low-power computing based on commodity technology
- MegaProto/Crusoe (version 1)
  - (TM5800@933MHz ＋ 2 x 1GbE) x 16 ＝14.9GFlops@300W (50MFlops/W)
  - 1.4-2.8 x dual-Xeon (IS,MG,EP,FT)
  - March, 2004 : 2 Unit (32 PE)
  - good performance/power
- MegaProto/Efficeon (version2)
  - (TM8820@1.0GHz ＋ 2 x 1GbE) x 16
  - June, 2005 : 20 Unit (320 PE)

MegaProto/Efficeon (version 2)

- delivered yesterday!