The IBM-DOME 64bit µServer Demonstrator: Findings, Status And Outlook

Ronald P. Luijten – Data Motion Architect
lui@zurich.ibm.com
IBM Research - Zurich
8 April 2014

DISCLAIMER: This presentation is entirely Ronald's view and not necessarily that of IBM.

ASTRON

IBM

Compute is free – data is not

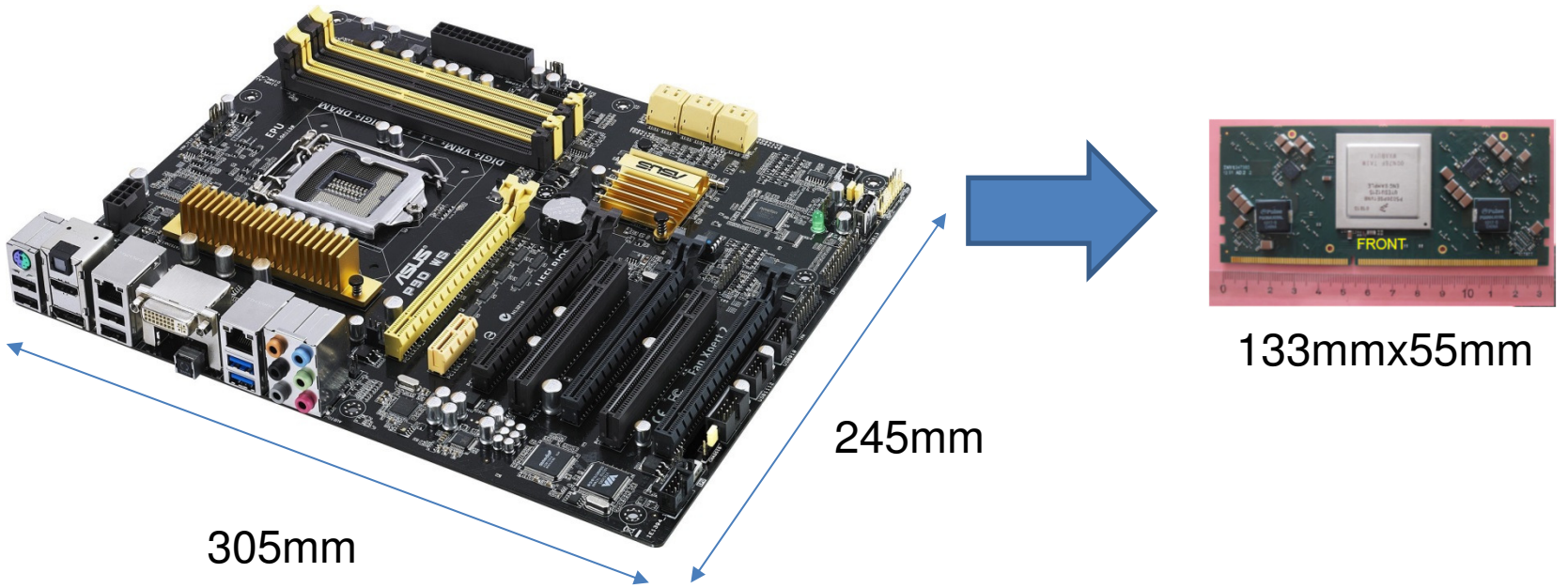
Ronald P. Luijten – Data Motion Architect
lui@zurich.ibm.com
IBM Research - Zurich
8 April 2014

DISCLAIMER: This presentation is entirely Ronald's view and not necessarily that of IBM.

Definition

µServer:

The integration of an entire server node motherboard* into a *single microchip* except DRAM, Nor-boot flash and power conversion logic.

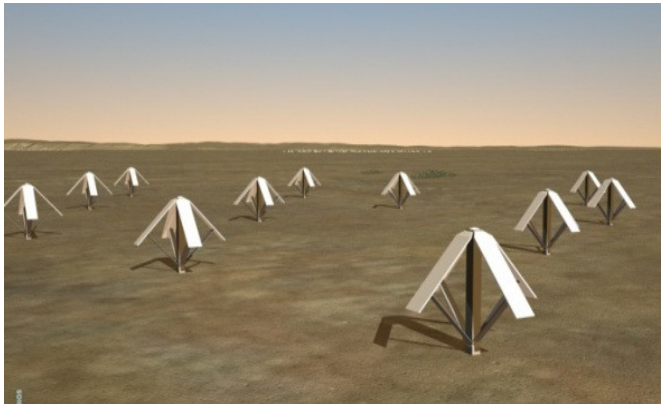


* no graphics

SKA (Square Kilometer Array) to measure Big Bang

SKA: Largest Radio-astronomy antenna

→ Big data on Steroids



Up to 2 Million+ Antenna's
What does this mean?

Prelim. Spec. SKA, R.T. Schilizzi et al. 2007 / Chr. Broekema

~ 10 Pb/s

86'400 sec/day

10..14 ExaByte/day

Central Signal Proc

Correlator / Beamformer

AA 250x 16Tb/s = 4 Pb/s

AA 250x 16Tb/s = 4 Pb/s

Dishes 3000x 60Gb/s = 180 Tb/s

?

switch

1 – 500 Tb/s

Major cycle

Visibility processors

2 – 6 orders magnitude data reduction

Minimum required Flops

200 PFlops – 2.5 EFlops

Science Data Proc

Image formation

Science analysis, user interface& archive

O(10) GB/s

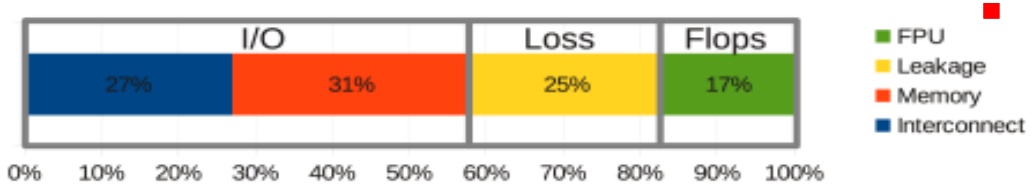
"Moore's Law"

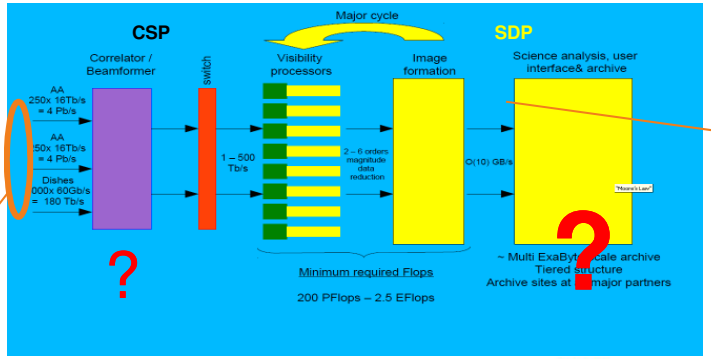
~ Multi ExaByte scale archive
Tiered structure
Archive sites at all major partners

~ 1 PB/Day.

Power distribution in a projected 2018 ExaFlop supercomputer

?





~ 10 Pb/s

86'400 sec/day

10..14 ExaByte/day

~ 1 PB/Day.

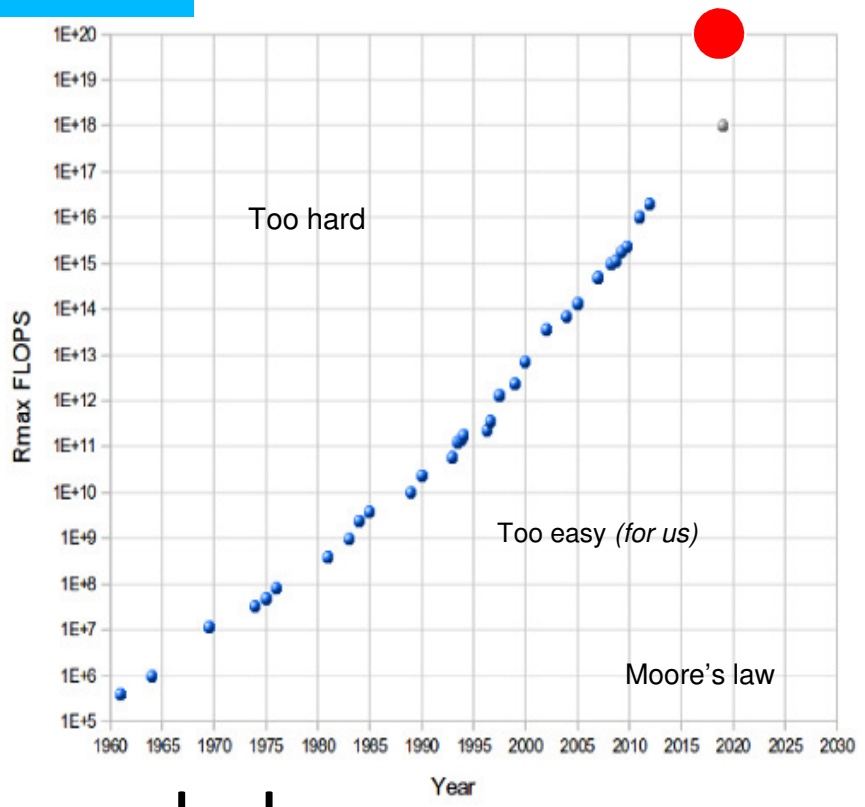
330 disks/day

120'000 disks/yr

Top-500 Supercomputing(11/2013)…. 0.3Watt/Gflop/s
→Today's industry focus is 1 Eflop @ 20MW. (2018)
➔(0.02 Gflop/s)

➔Most recent data from SKA:
- ➔CSP….max. power 7.5MW
- ➔SDP….max. power 1 MW
- ➔Latest need for SKA – 4 Exaflop (SKA1 - Mid)
- ➔ 1.2GW…80MW

Factor 80-1200



→multiple breakthroughs needed

IBM / ASTRON DOME project
Technology roadmap development

•Sustainable (Green) Computing

•Nanophotonics

•Data & Streaming

•User Platform

•System Analysis

•Algorithms & Machines

-Student projects
-Events
-Research Collaboration

•Computing

-Microservers
-Accelerators

•Transport

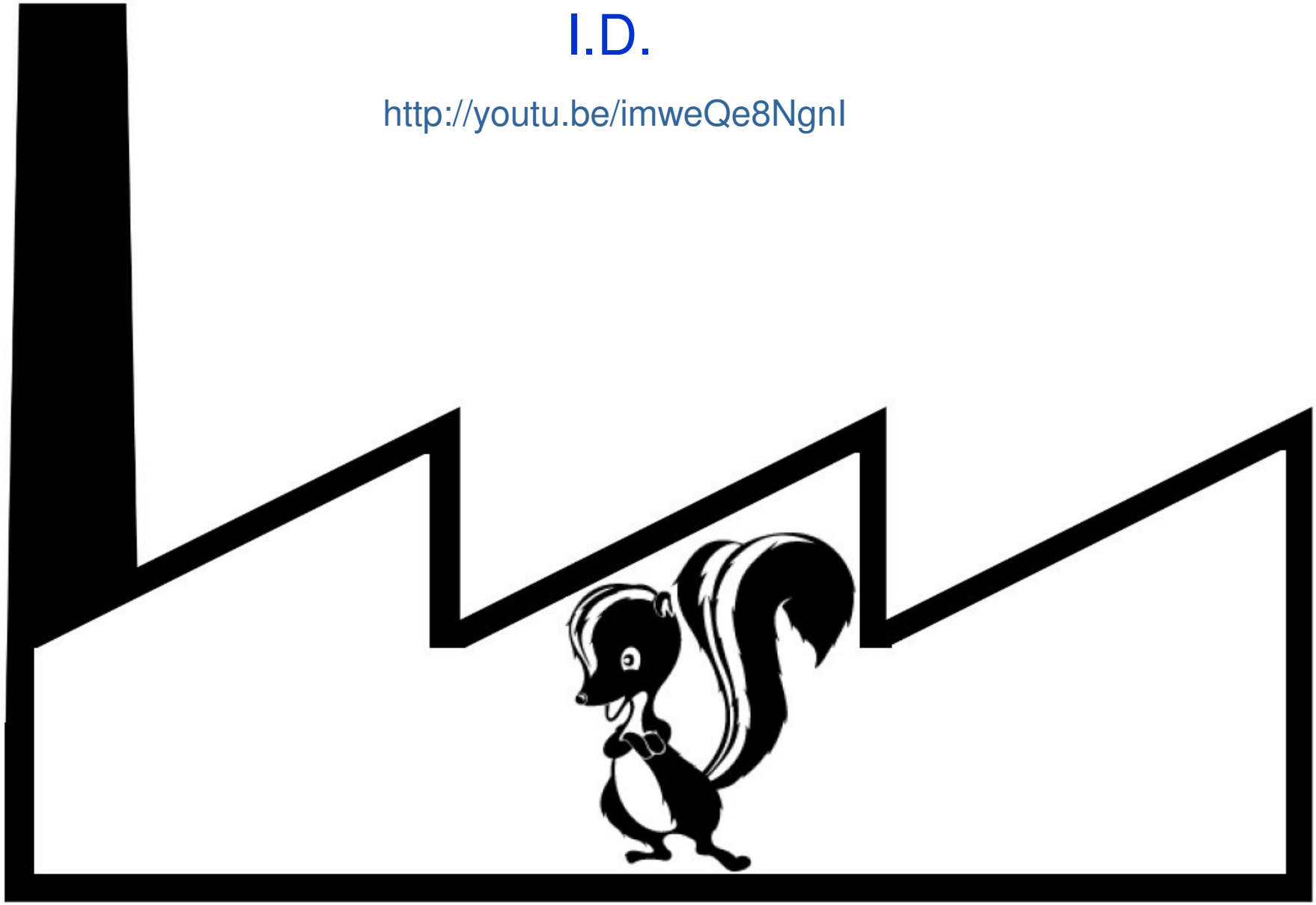
-Nanophotonics
-Real Time Communications
-Compressive Sampling

•Storage

-Access Patterns

I.D.

http://youtu.be/imweQe8NgnI



IBM DOME µServer Motivation & Objectives

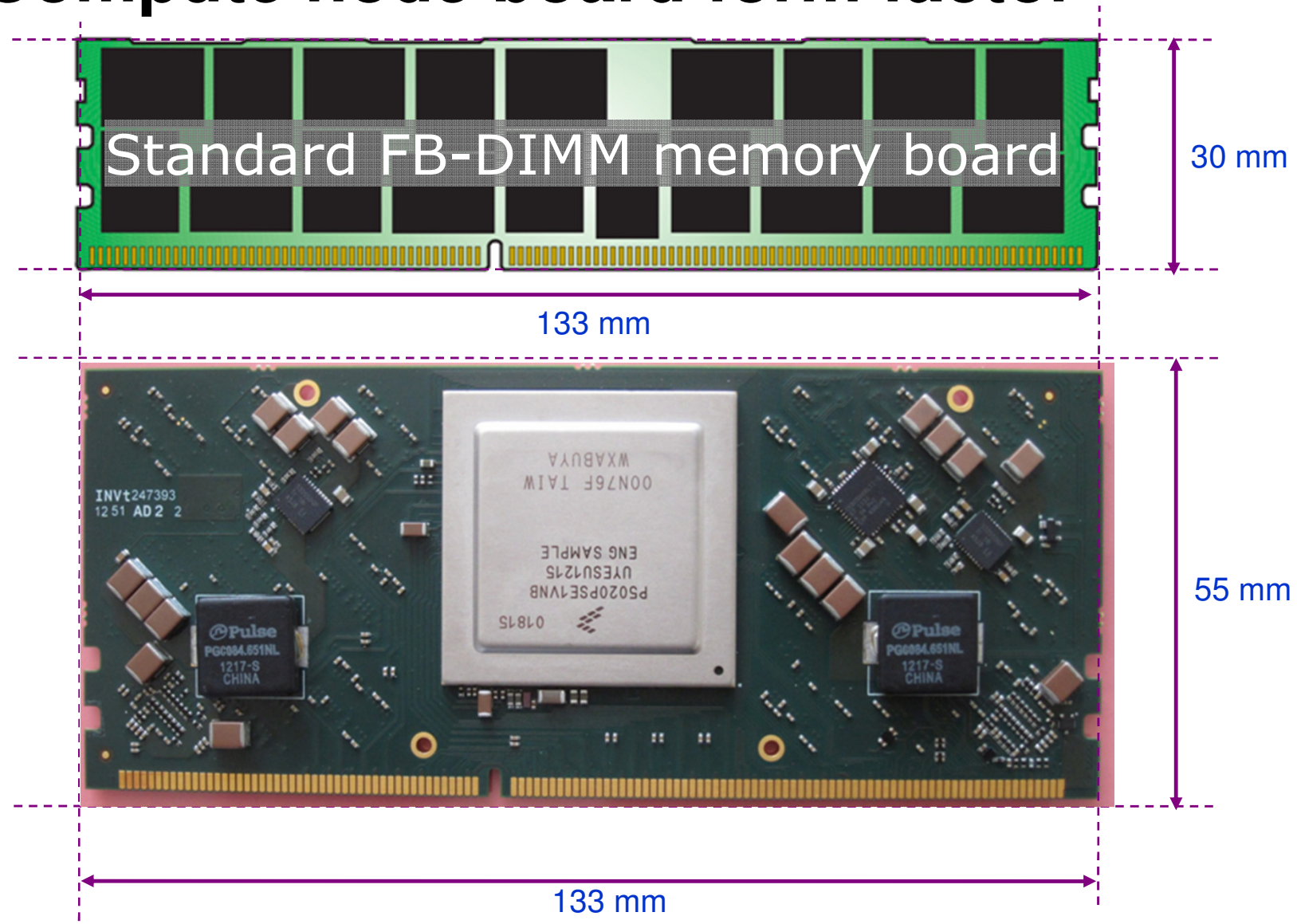
- **Create *the* worlds highest density 64 bit µ-server drawer**
  - Useful for both SKA radio-astronomy and IBM future business
    - Platform for Business Analytics appliance pre-product research
    - "Datacenter in-a-box"
  - Very high energy efficiency / very low cost (radioastronomers…)
  - Use commodity components only, HW + SW standards
  - Leverage 'free computing' paradigm
  - Enhance with 'Value Add': packaging, system integration, …
  - Density and speed of light

- **Most efficient cooling using IBM technology (ref: SuperMUC TOP500 machine)**

- **Must be true 64 bit to enable business applications**
- Must run server class OS (SLES11 or RHEL6, or equivalent)
  - Precluded ARM (64-bit Silicon was not available)
  - PPC64 is available in SoC from FSL since 2011
  - (I am poor – no $$$ for my own SoC…)

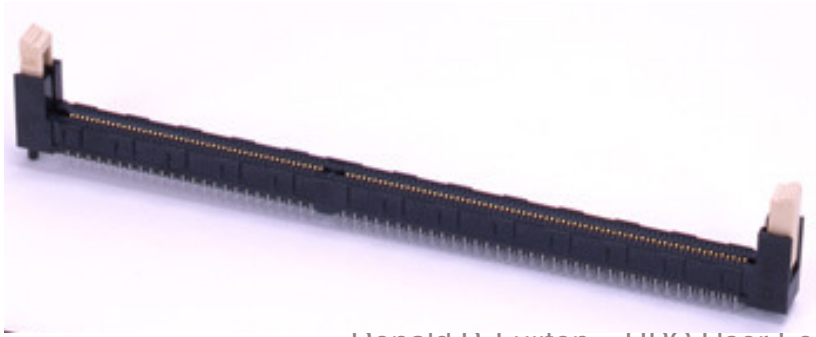
- **This is a research project – capability demonstrator only**

Compute node board form factor



Compute node processor options

FSL SoC parts	P5040	T4240
CPU GHz	2.2	1.8
CPUs	4 cores, 1 thread per core	12 cores, 2 threads per core
Primary cache	32 KB I + 32 KB D per core	32 KB I + 32 KB D per core
Secondary cache	512 KB I+D	2 MB per 4 CPUs
L3 cache	1 MB on chip	1.5 MB on chip
Memory	2 x 2 GB, DDR3/L3, ECC	3 x 2 GB, DDR3/L3, ECC
core	e5500, ppc64	e6500, ppc64
	1 DP FP unit per core	1 DP FP unit per core 128 bit SP altivec unit per core
node	45nm	28nm
TDP	55W	60W



T4240 DIMM connector:

- 2 times SATA
- 4 times 10 Gigabit ethernet
- SD card interface
- USB interface
- Some power supplies

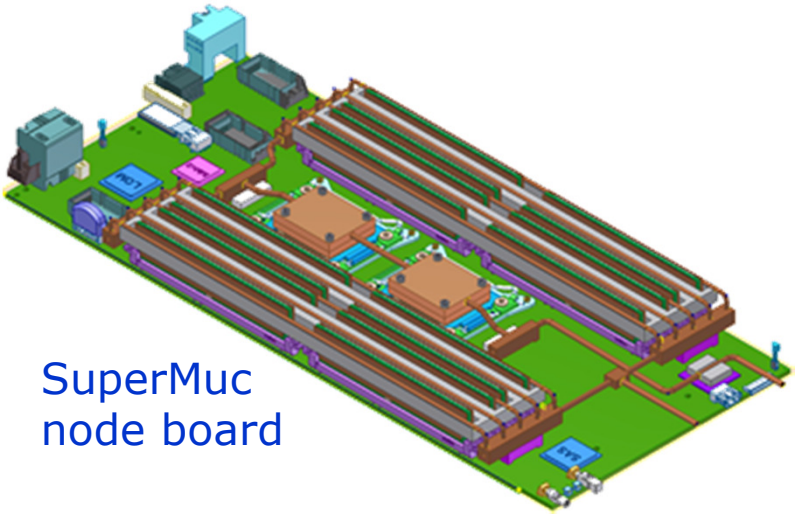
Hot Water Cooling

Most Energy Efficient solution:

- Low PUE possible (<=1.1) – Green IT
- 40% less energy consumption compared to air-cooled systems
- 90% of waste heat can be reused (CO_2 neutral according Kyoto protocol)
- Allows very high density
- Less thermal cycling - improved reliability
- Lower T_j reduces leakage current – further saving energy

SuperMUC HPC machine at LRZ in Germany demonstrates ZRL hot water cooling

- No 4 on June 2012 TOP500 HPC list

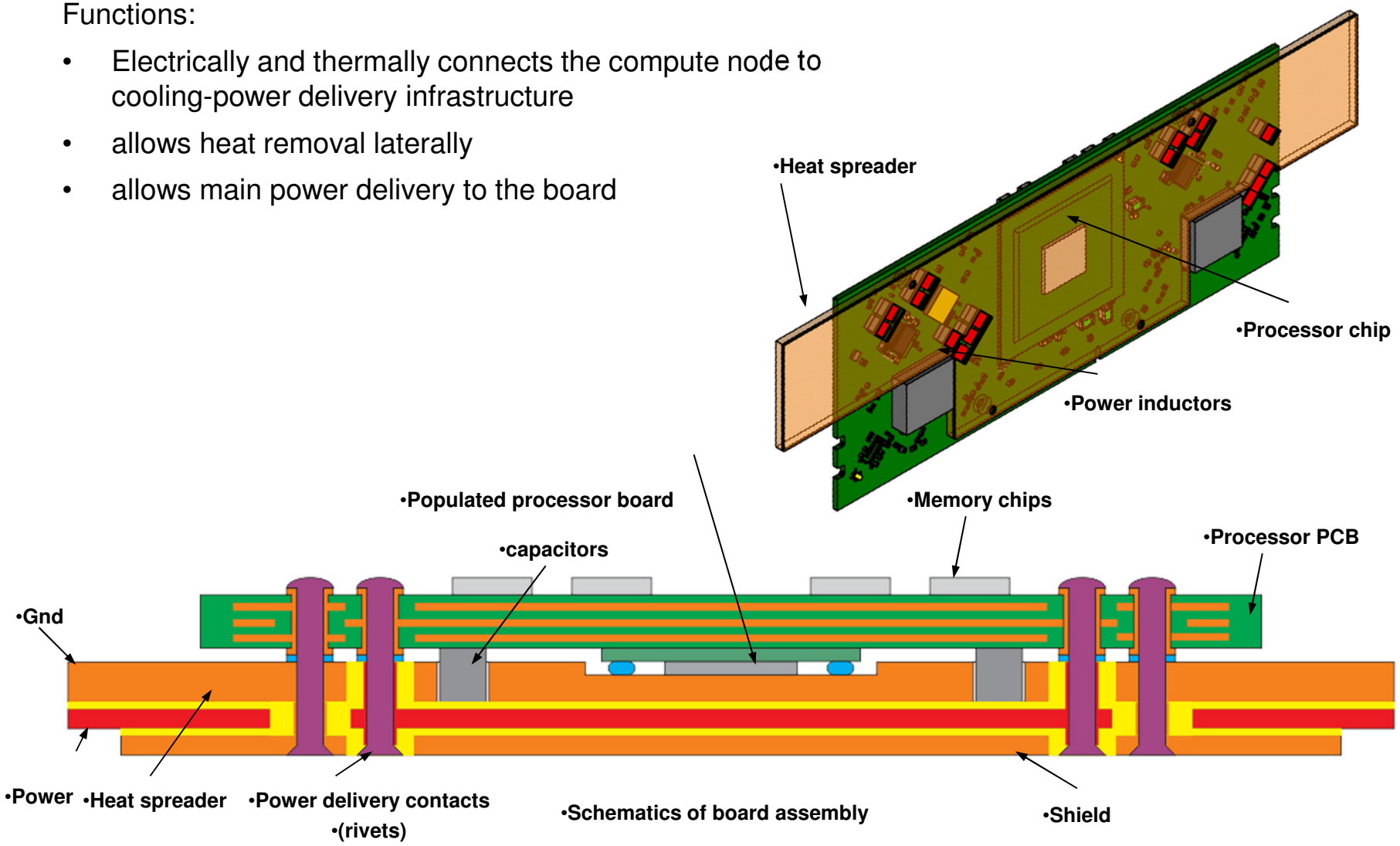


SuperMuc
node board

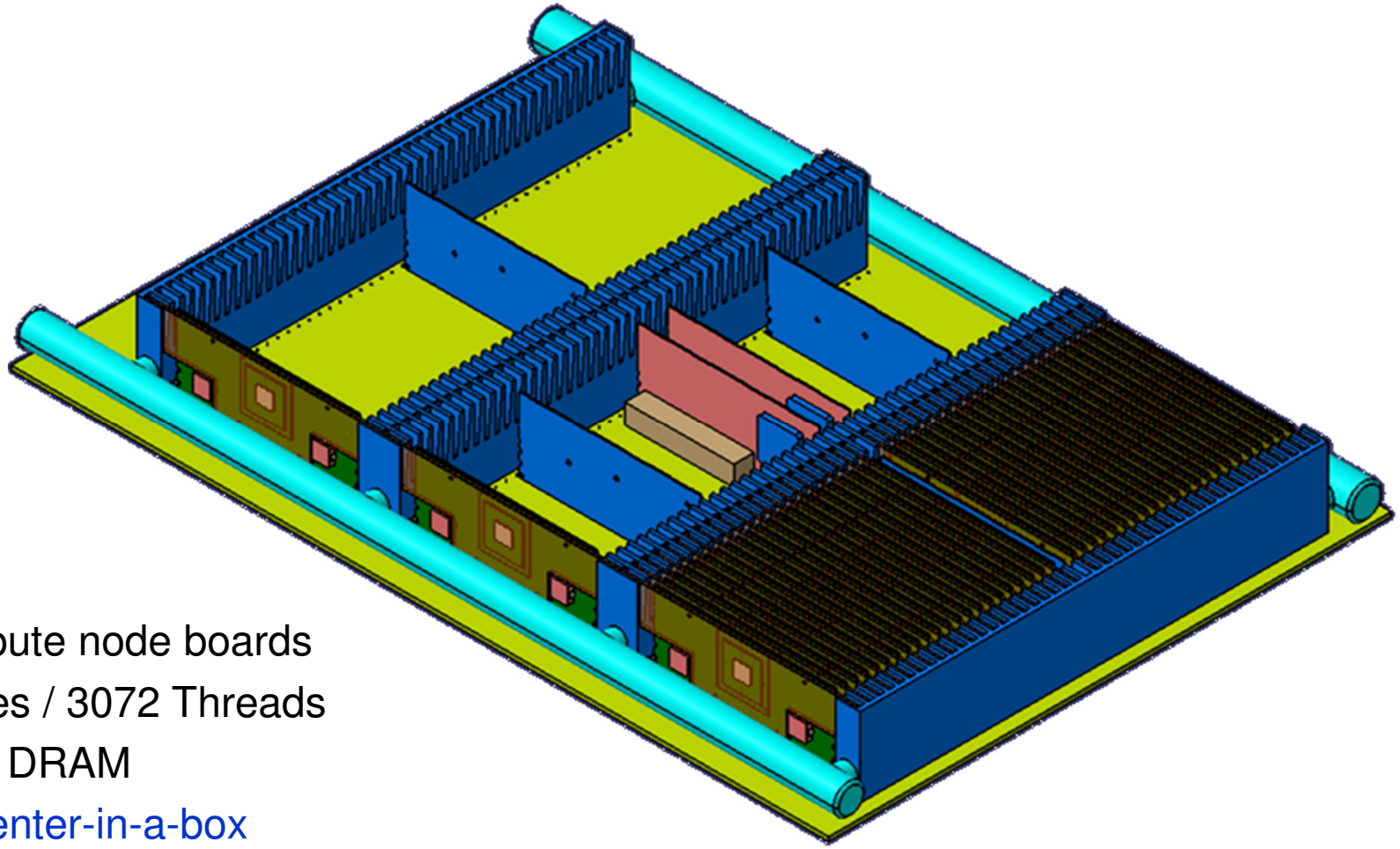
Compute node heat spreader

Functions:

- Electrically and thermally connects the compute node to cooling-power delivery infrastructure
- allows heat removal laterally
- allows main power delivery to the board



19” 2U Chassis with Combined Cooling and Power



128 compute node boards

1536 cores / 3072 Threads

3 or 6 TB DRAM

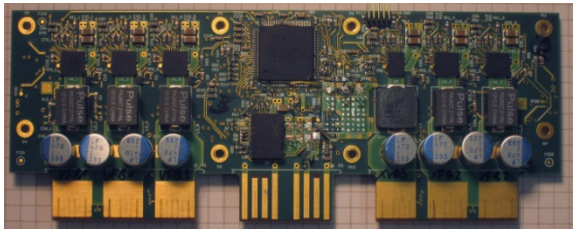
→ Datacenter-in-a-box

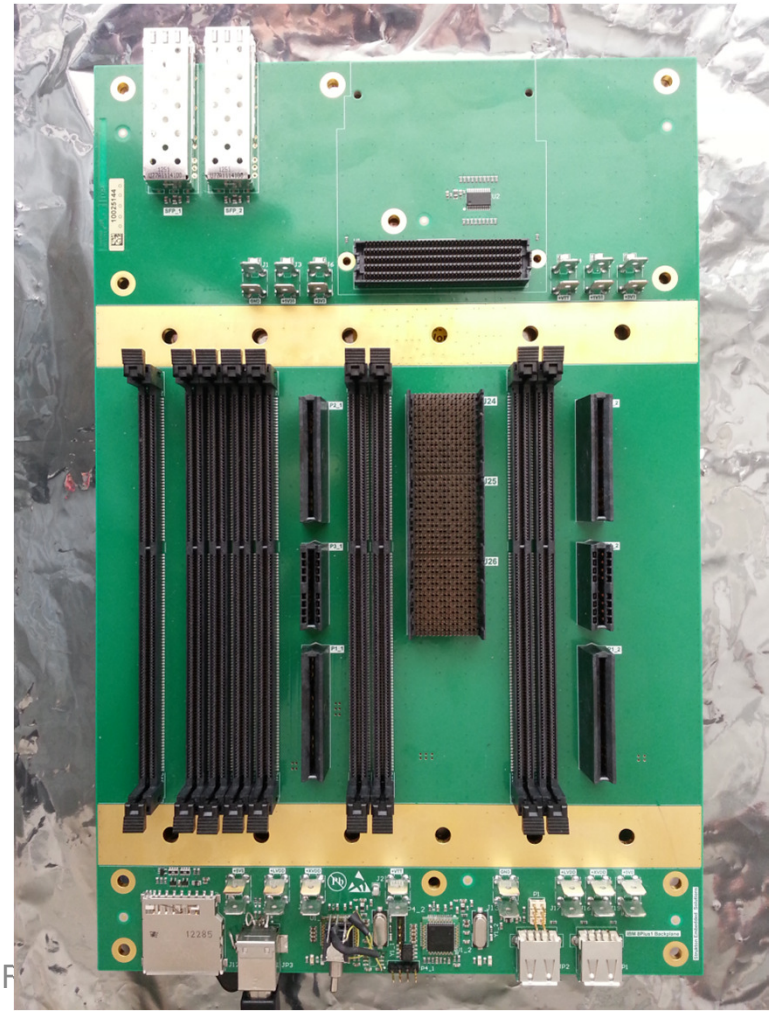
Lessons learnt

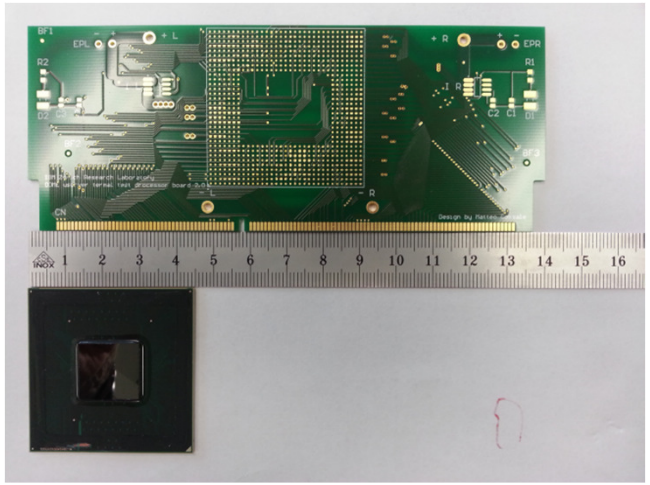
- I underestimated the effort to build a compute node board using an SoC
  - 2000 page manuals, 300 page manuals, more manuals
  - Reset
  - Embedded area not as easy as PC area…. We are spoilt.
- I underestimated the software effort to get it to boot linux
  - Yocto, Uboot, tiny loader, PBL programming, kernel configuration, dtb, cross-compilation etc.
- I overestimated market readiness for µServers
  - Ie. It came / comes much later
  - (net: we are well positioned timewise)
- Linux / SW Ecosystem
- Performance Benchmarks
  - Stream, specbench, performance / energy scaling

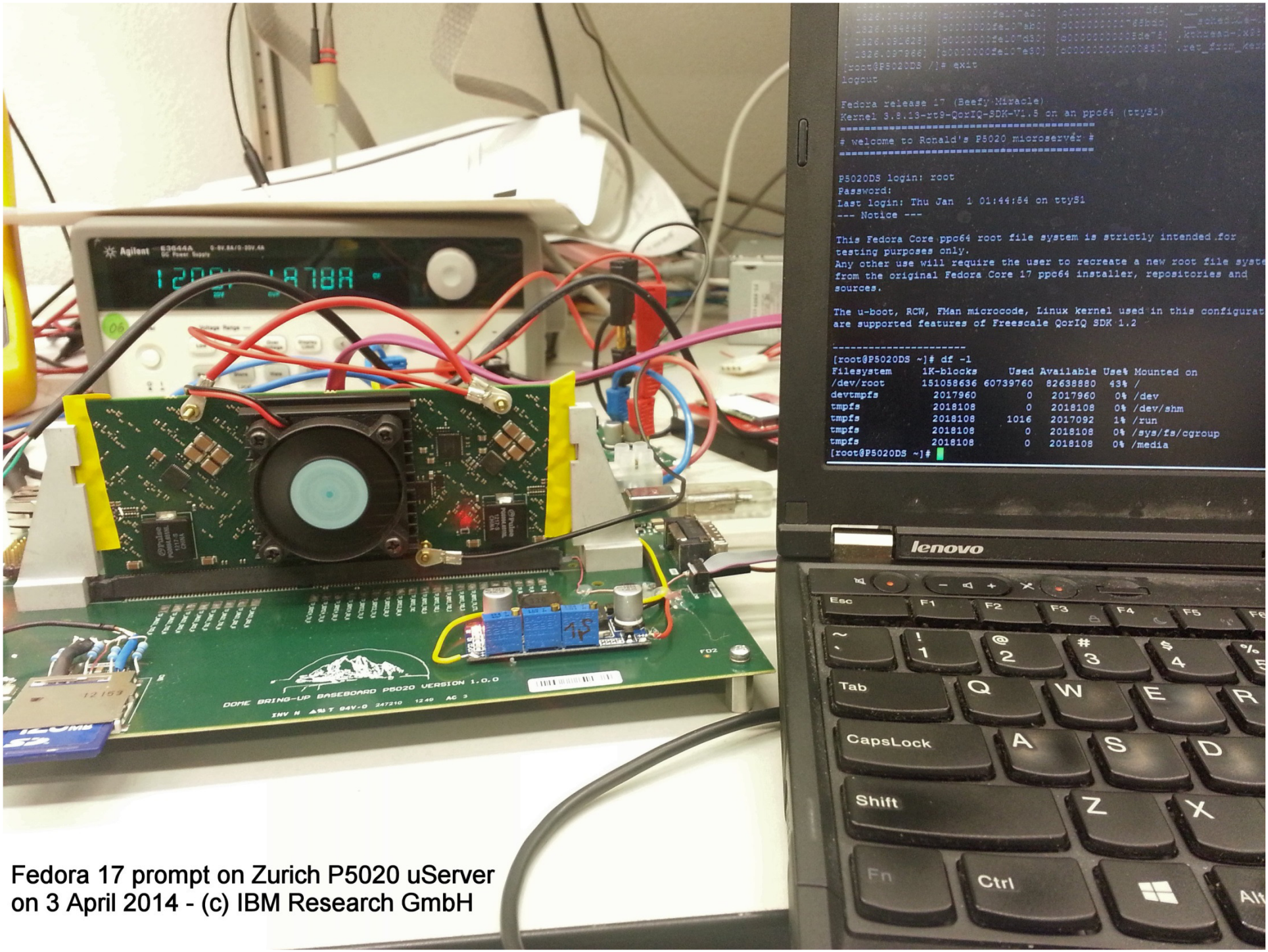
Status (3 april 2014)

- Rev 2 P5020/P5040 board in bringup
  - Uboot is running, Sata works, booted Fedora 17, ppc64
- Rev 1 T4240 board being populated
  - At our lab in 7 weeks
- Power module being debugged
- Multinode carrier board in bringup
  - 8 P50x0 compute nodes
  - 10Gbps Ethernet switch module
  - Power module
  - Integrated water cooling
- Water Cooling Thermal Test Vehicle in bringup



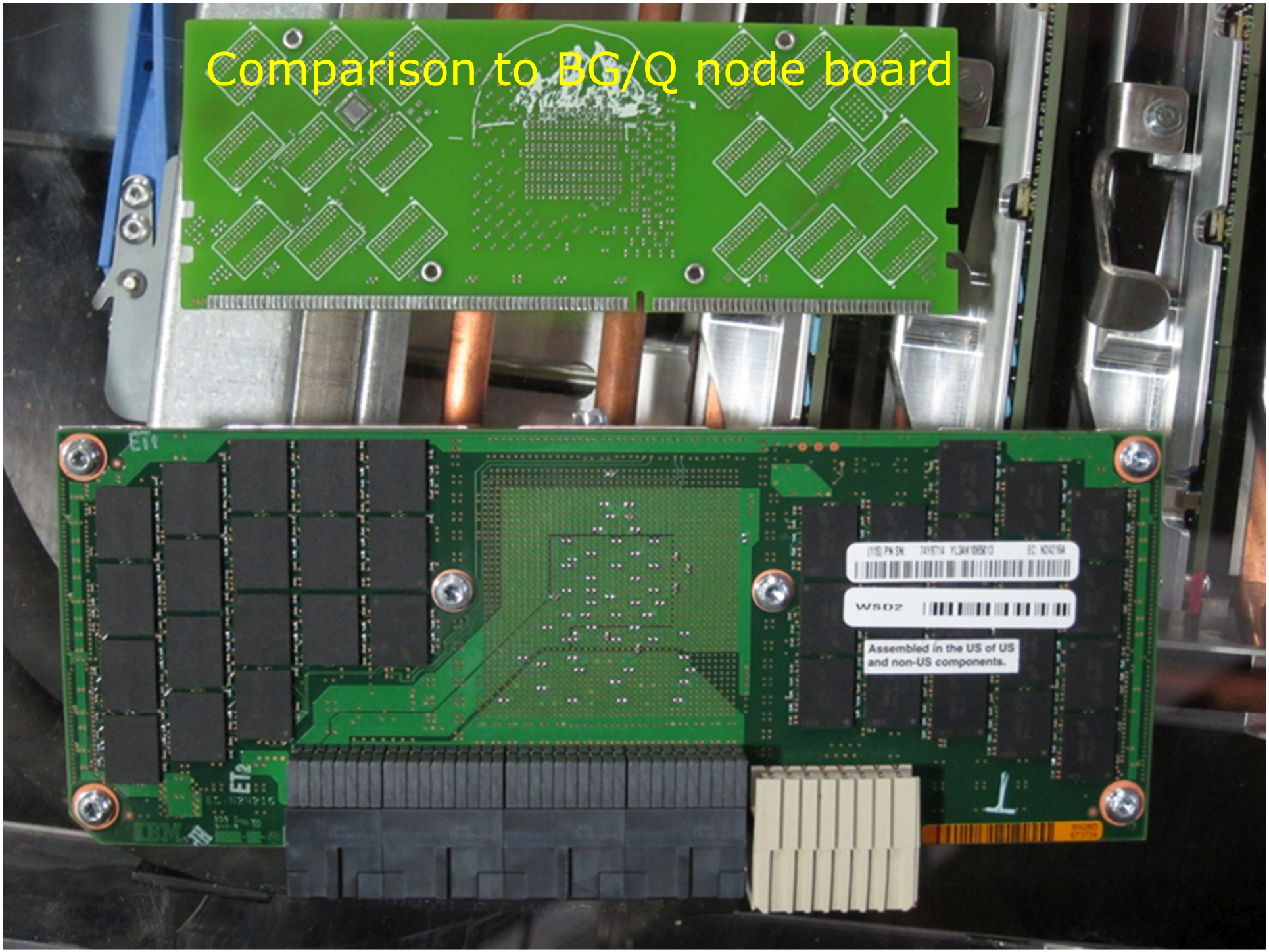




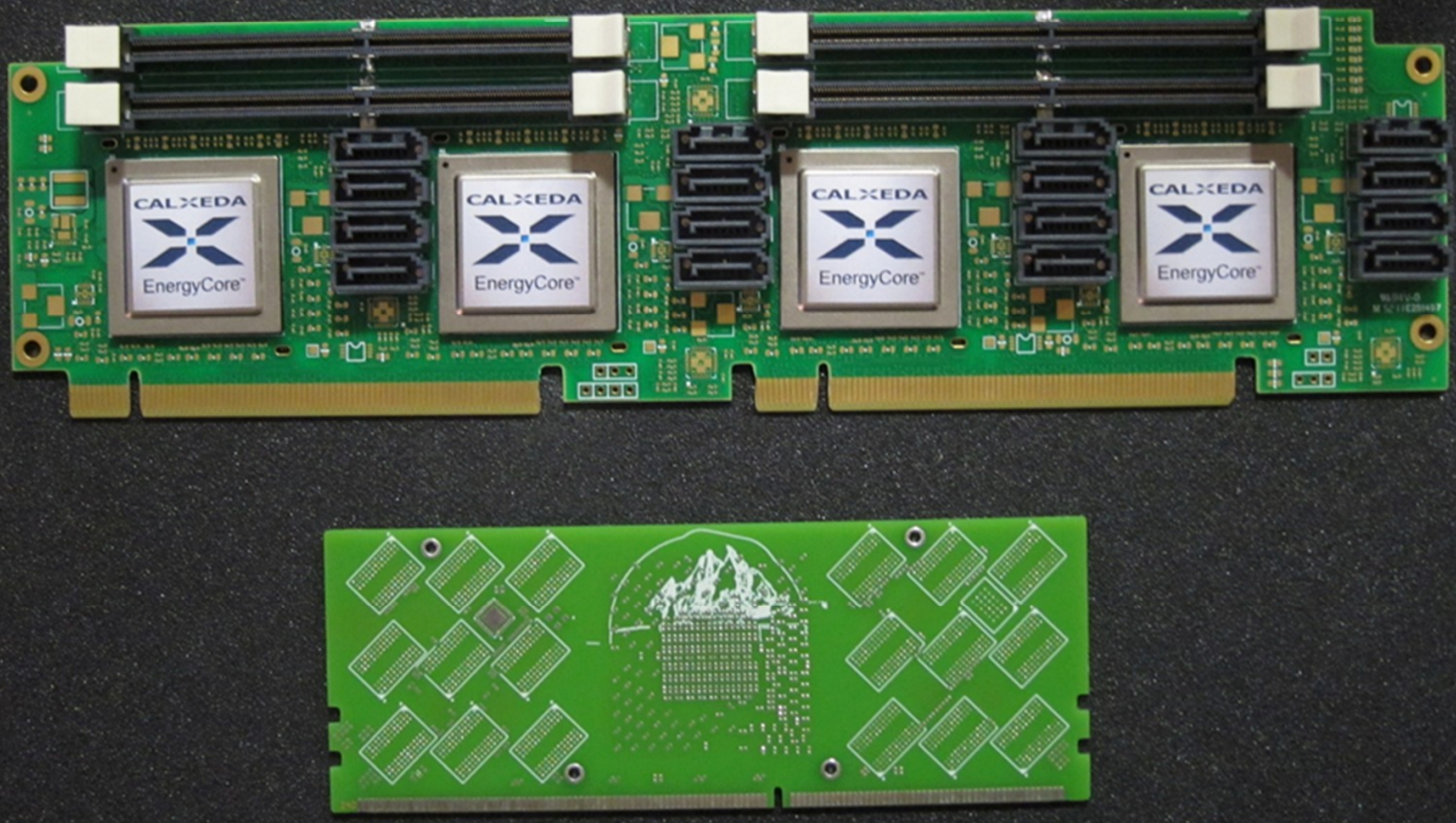



```
Fedora release 17 (Beefy Miracle)
Kernel 3.8.13-rt9-QorIQ-SDK-V1.5 on an ppc64 (ttyS1)
=========================================
# Welcome to Ronald's P5020 microserver #
=========================================

P5020DS login: root
Password:
Last login: Thu Jan  1 01:44:54 on ttyS1
--- Notice ---

This Fedora Core ppc64 root file system is strictly intended for
testing purposes only.
Any other use will require the user to recreate a new root file syste
from the original Fedora Core 17 ppc64 installer, repositories and
sources.

The u-boot, RCW, FMan microcode, Linux kernel used in this configurat
are supported features of Freescale QorIQ SDK 1.2

----------------------
[root@P5020DS ~]# df -l
Filesystem     1K-blocks      Used Available Use% Mounted on
/dev/root      151058636 60739760  82638880  43% /
devtmpfs         2017960         0   2017960   0% /dev
tmpfs            2018108         0   2018108   0% /dev/shm
tmpfs            2018108      1016   2017092   1% /run
tmpfs            2018108         0   2018108   0% /sys/fs/cgroup
tmpfs            2018108         0   2018108   0% /media
[root@P5020DS ~]#
```

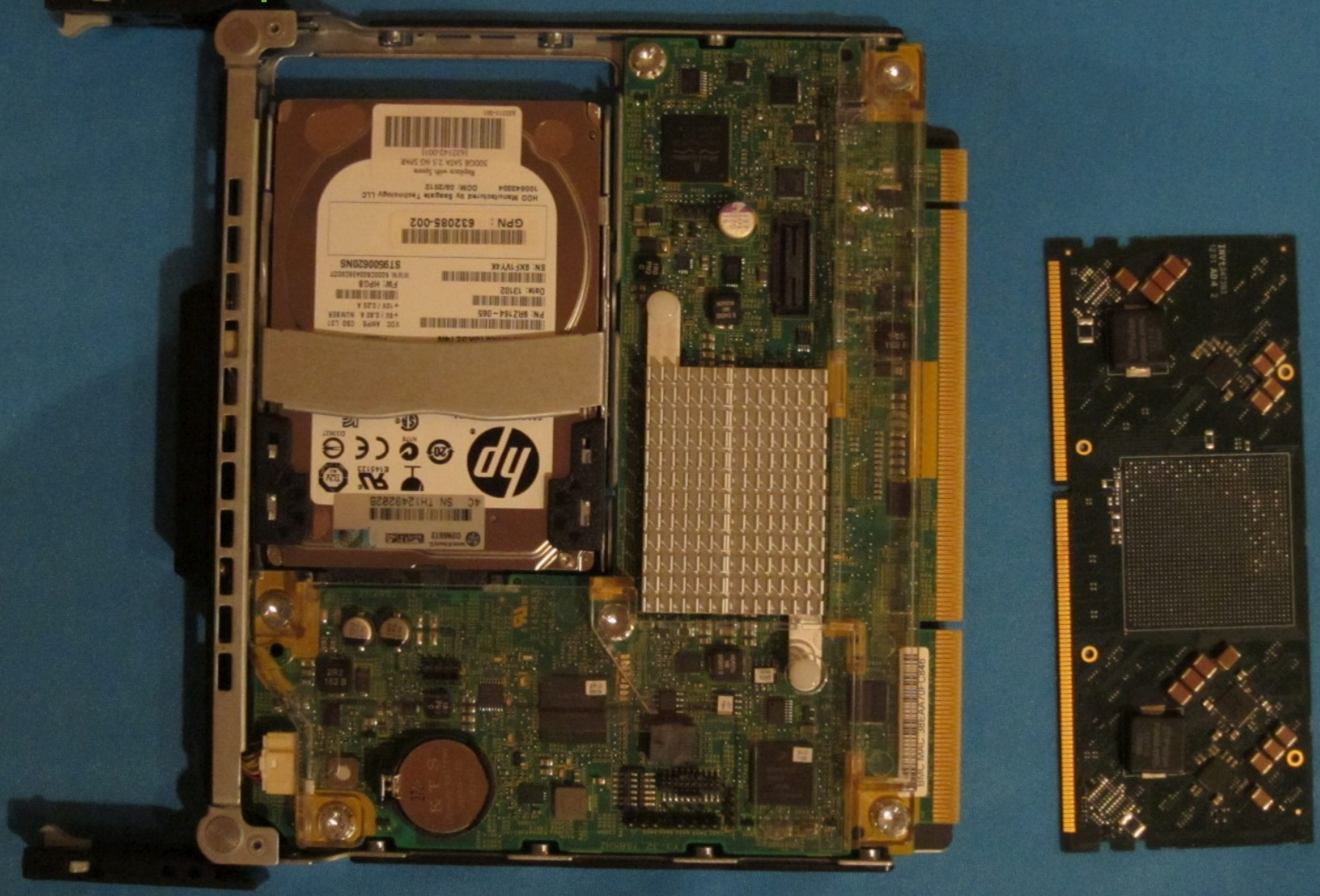
Fedora 17 prompt on Zurich P5020 uServer
on 3 April 2014 - (c) IBM Research GmbH


Comparison to BG/Q node board
ET1
ET2
IBM
W8D2
Assembled in the US of US
and non-US components.

Comparison to Calxeda node board



Comparison to Moonshot node board



Acknowledgements

This work is the results of many *people*

- Peter v. Ackeren, FSL
- Ed Swarthout, FSL Austin
- Dac Pham, FSL Austin
- Yvonne Chan, IBM Toronto
- Andreas Doering, IBM ZRL
- Tom Wilson, IBM Armonk
- Alessandro Curioni, IBM ZRL
- Stephan Paredes, IBM ZRL
- James Nigel, FSL
- Gary Streber, FSL
- Patricia Sagmeister, IBM ZRL
- Boris Bialek, IBM Toronto
- Marco de Vos, Astron NL
- Hillery Hunter, IBM WRL
- Vipin Patel, IBM Fishkill
- And many more remain unnamed….

Companies: FSL Austin, Belgium & Germany; IBM worldwide; Transfer - NL

Questions???

µServer website: www.swissdutch.ch

Published Conference Papers

- **“Parallelism and Data Movement Characterization of contemporary Application Classes** ”, Victoria Caparros Cabezas, Phillip Stanley-Marbell, ACM SPAA 2011, June 2011
- “**Quantitative Analysis of the Berkeley Dwarfs' Parallelism and Data Movement Properties**”, Victoria Caparros Cabezas, Phillip Stanley-marbell, ACM CF 2011, May 2011
- “**Performance, Power, and Thermal Analysis of Low-Power Processors for Scale-Out Systems**”, Phillip Stanley-Marbell, Victoria Caparros Cabezas, IEEE HPPAC 2011, May 2011
- “**Pinned to the Walls—Impact of Packaging and Application Properties on the Memory and Power Walls**”, Phillip Stanley-Marbell, Victoria Caparros Cabezas, Ronald P. Luijten, IEEE ISLPED 2011, Aug 2011.
- “**The DOME embedded 64 bit microserver demonstrator**”, R. Luijten and A. Doering, ICICDT 2013, Pavia, Italy, May 2013