SKA, DOME & ASTRON project - µServer

Ronald P. Luijten – Data Motion Architect
lui@zurich.ibm.com
IBM Research - Zurich
16 July 2015

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

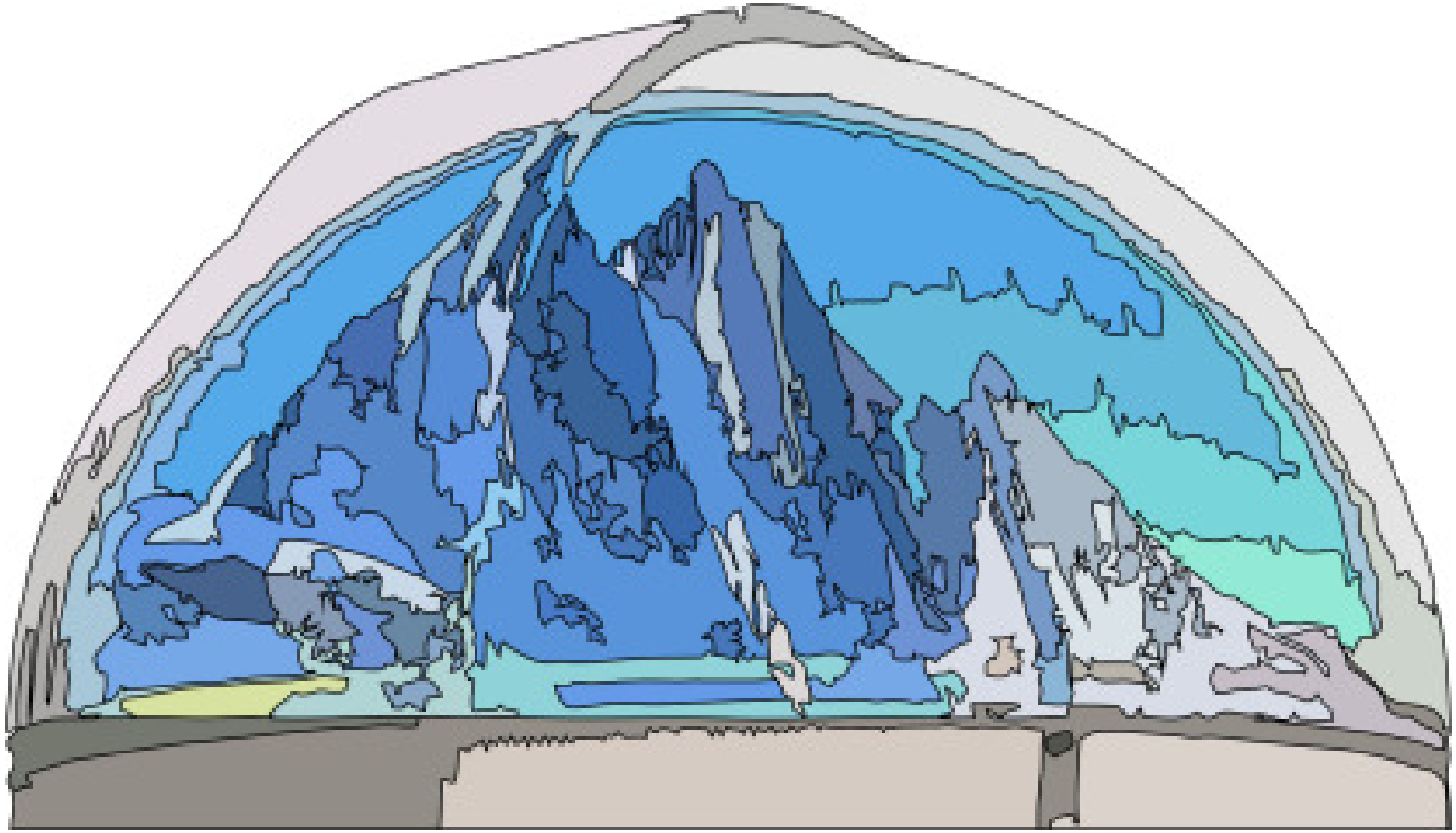
ASTRON

IBM

COMPUTE is FREE – DATA is NOT

Ronald P. Luijten – Data Motion Architect
lui@zurich.ibm.com
IBM Research - Zurich
16 July 2015

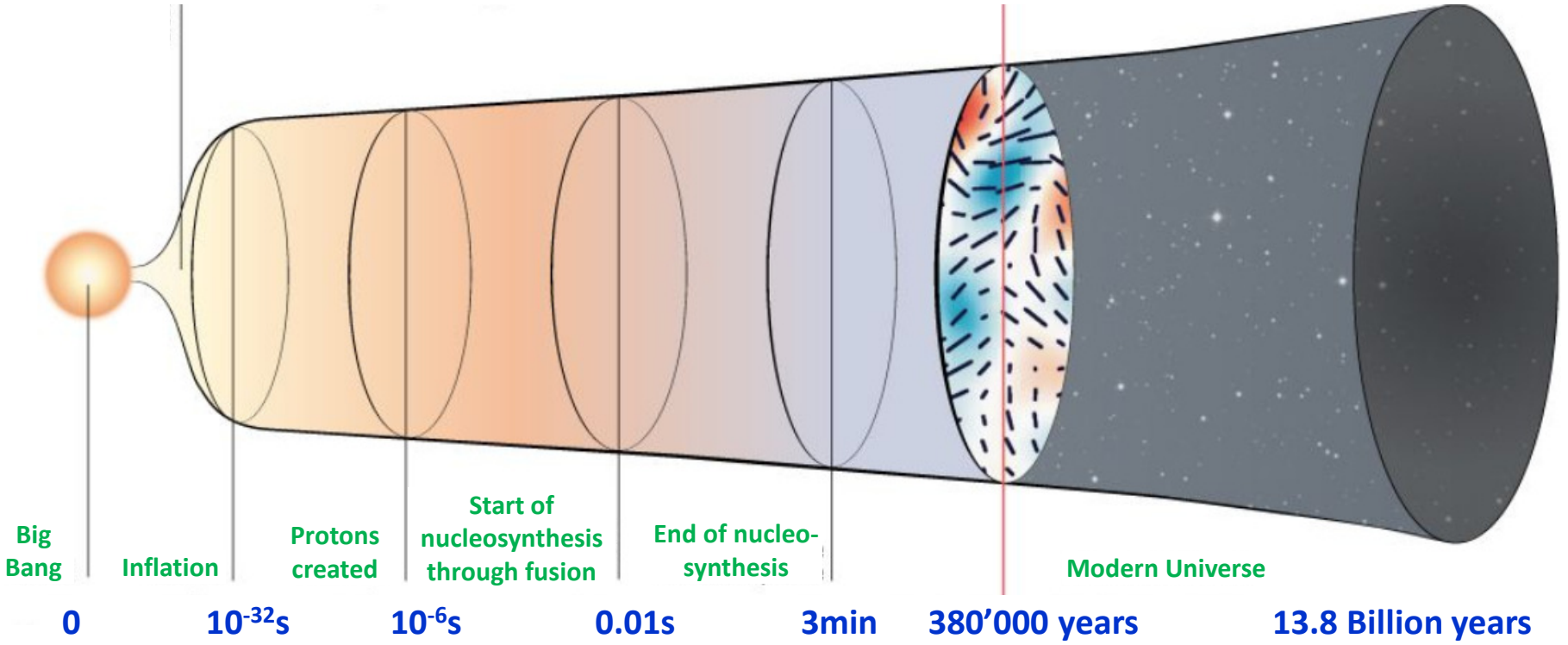
DISCLAIMER: This presentation is entirely Ronald's view and not necessarily that of IBM.



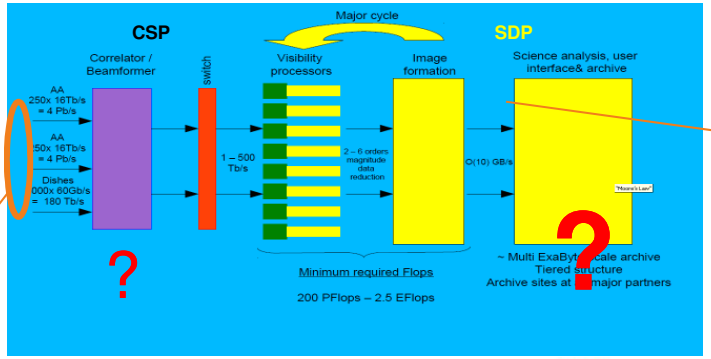
DOME:

- ppp Astron, IBM, Dutch gvt
- 20MEur funding over 5 years
- Started feb 2012

SKA (Square Kilometer Array) to measure Big Bang



Picture source: NZZ march 2014



~ 10 Pb/s

86'400 sec/day

15 ExaByte/day

~ 1 PB/Day.

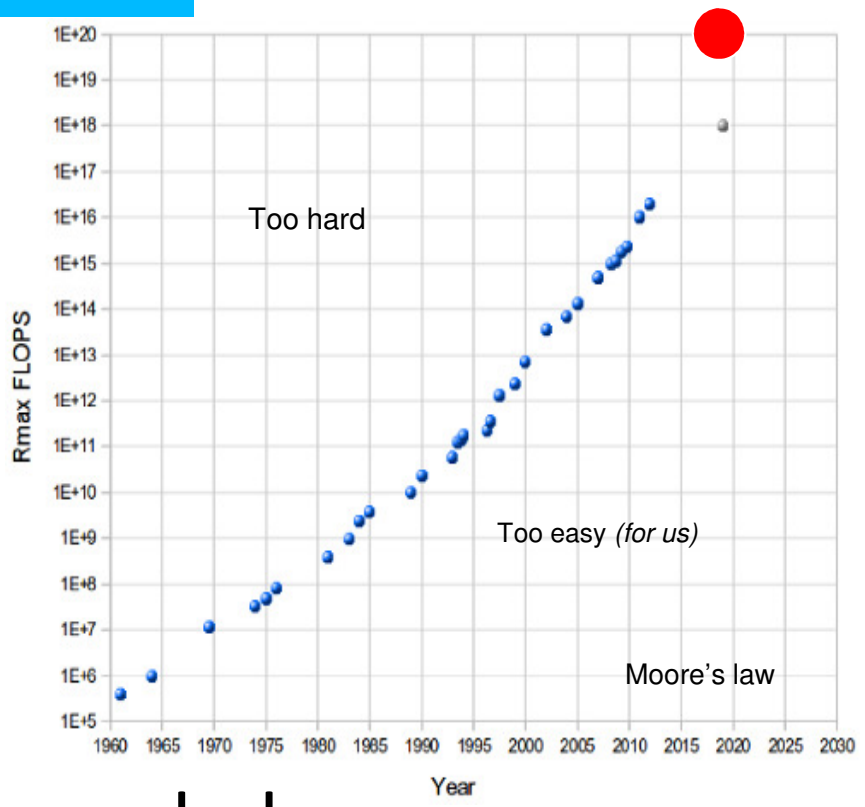
330 disks/day

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

DOME µServer Motivation & Objectives

•**Create *the* worlds highest density 64 bit µ-server drawer**
–Useful to evaluate both SKA radio-astronomy and IBM future business
–Platform for Business Analytics appliance pre-product research
–High energy efficiency / very low cost
–Commodity components, HW + SW standards based
–Leverage 'free computing' paradigm
–Enhance with 'Value Add': packaging, system integration, …
–Density and speed of light

•**Most efficient cooling using IBM technology**
(ref: SuperMUC June 2012 TOP500 machine)

•**Must be true 64 bit to enable business applications**

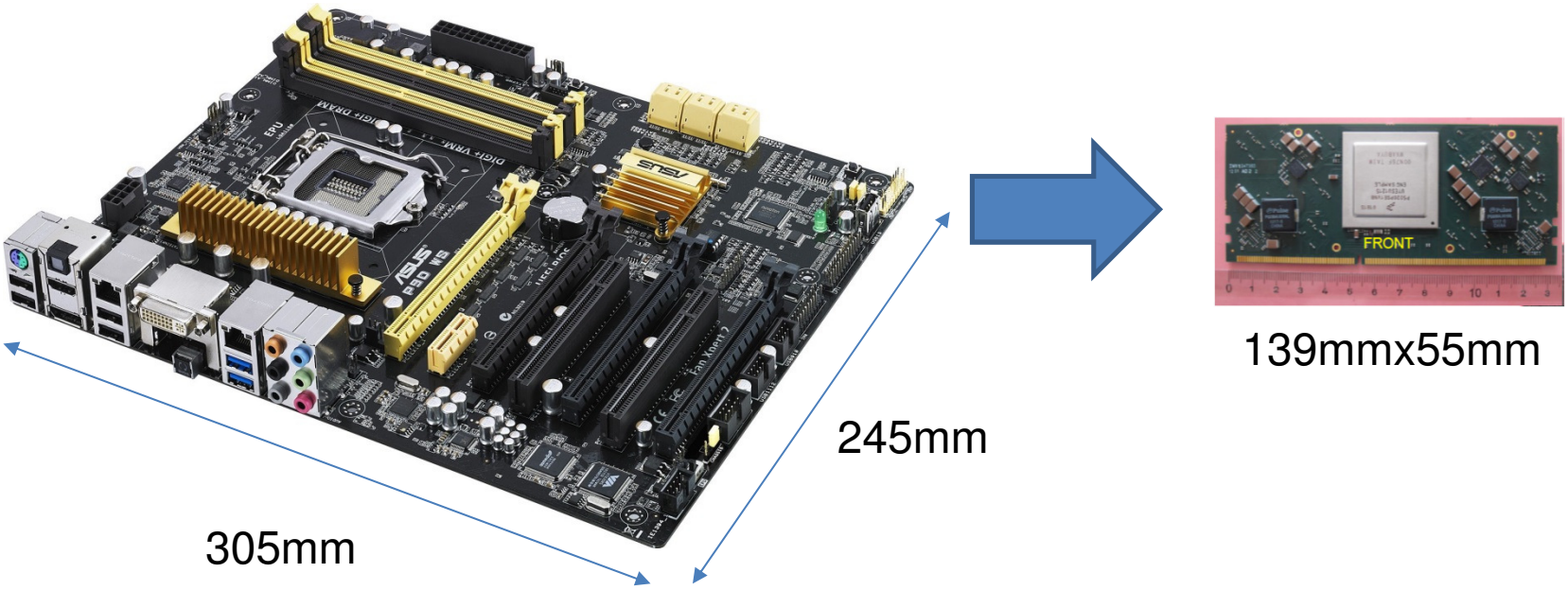
•**Must run server class OS (SLES11 or RHEL6, or equivalent)**
–Precluded ARM (64-bit Silicon was not available)
–PPC64 is available in SoC from FSL since 2011
–(no $$$ to build a new SoC…)

•**This is the DOME project capability demonstrator – not a product**

Definition

µServer:

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



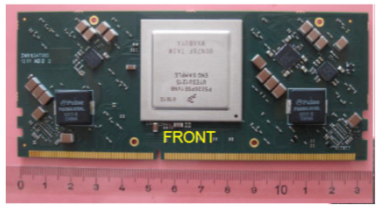
* no graphics

Definition

µServer:

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

This does NOT imply low performance!



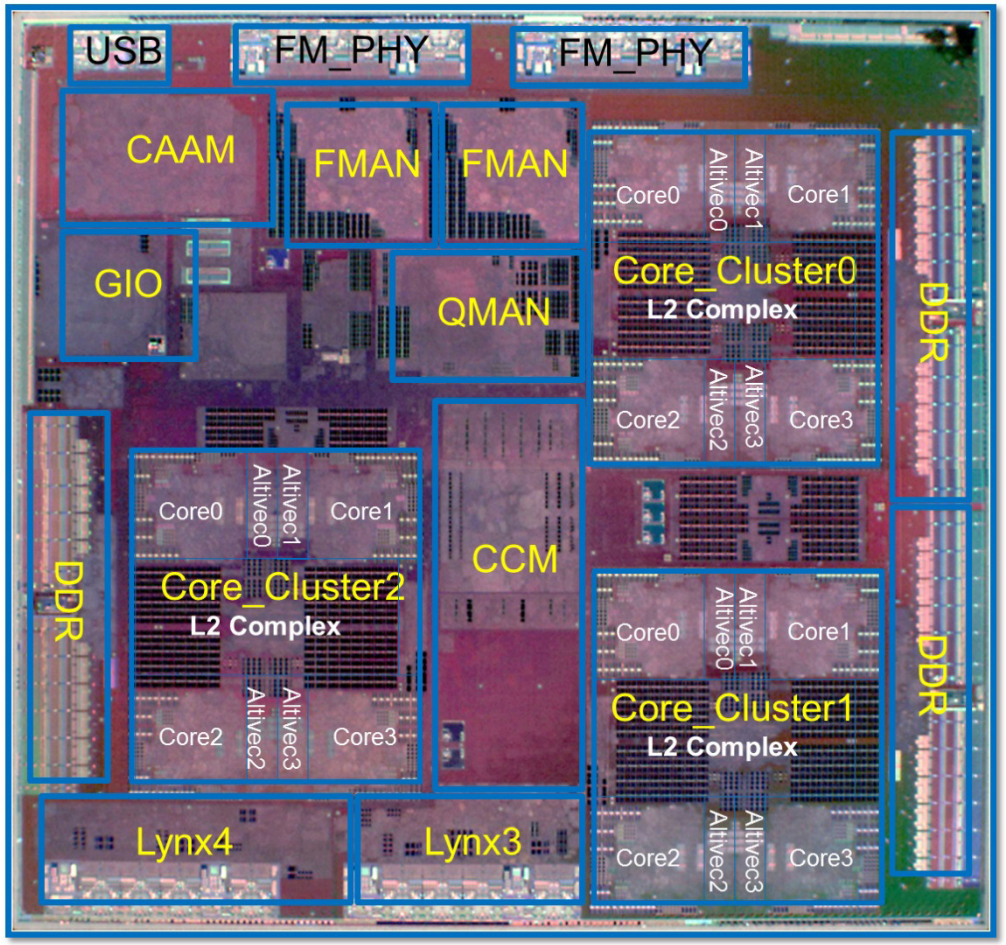
139mmx55mm

* no graphics

T4240 Chip Overview

12 core – fully dual threaded
1.8 GHz ppc64 (e6500)
12 DP-FPU; 12 128b Altivec
3 DDR3 channels at 1.86GT/s
3x 0.5MB L3 cache
4x 10GbE + 2x SATA
PCIe 3.0
HW packet acceleration
RegEx Pattern Match acc.
Crypto acceleration

28nm TSMC Bulk CMOS
239mm^2 - ~1.7B transistors
111Mbit SRAM, 6M FF



7 Power states (2 power gating)

T4240 Chip Overview

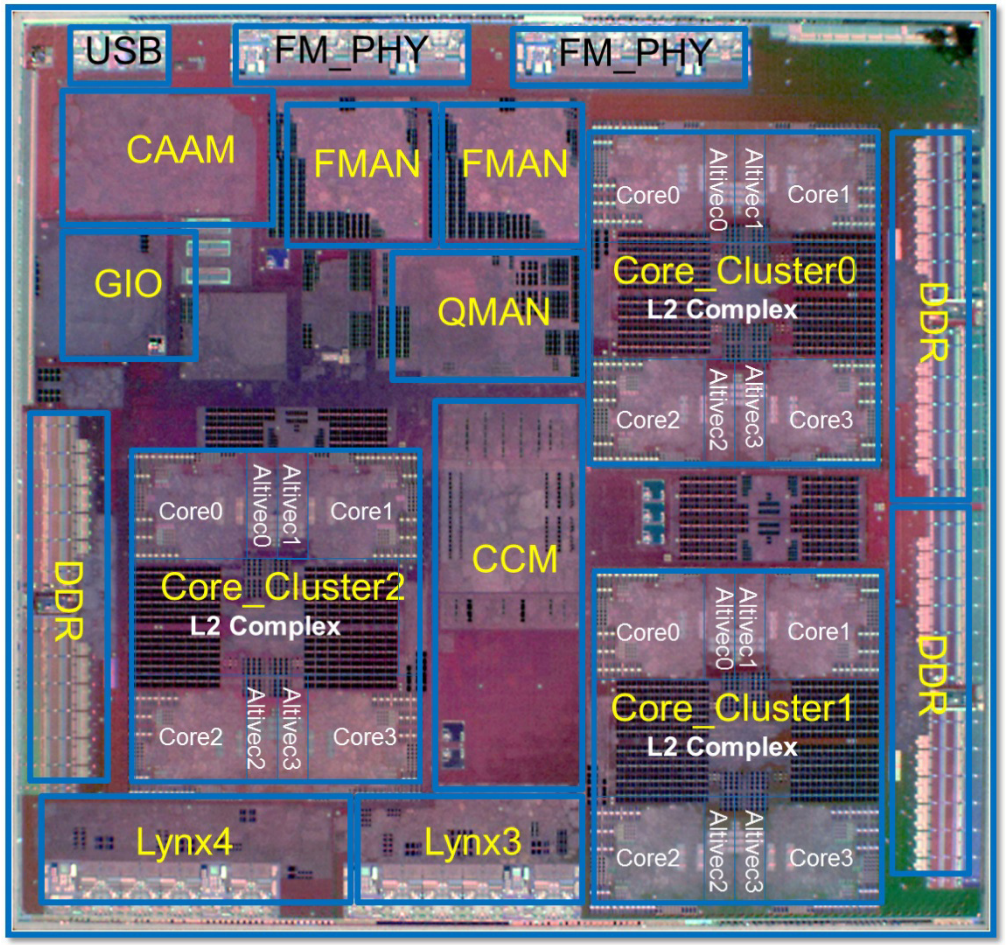
This is NOT the ideal part

However, a very good one

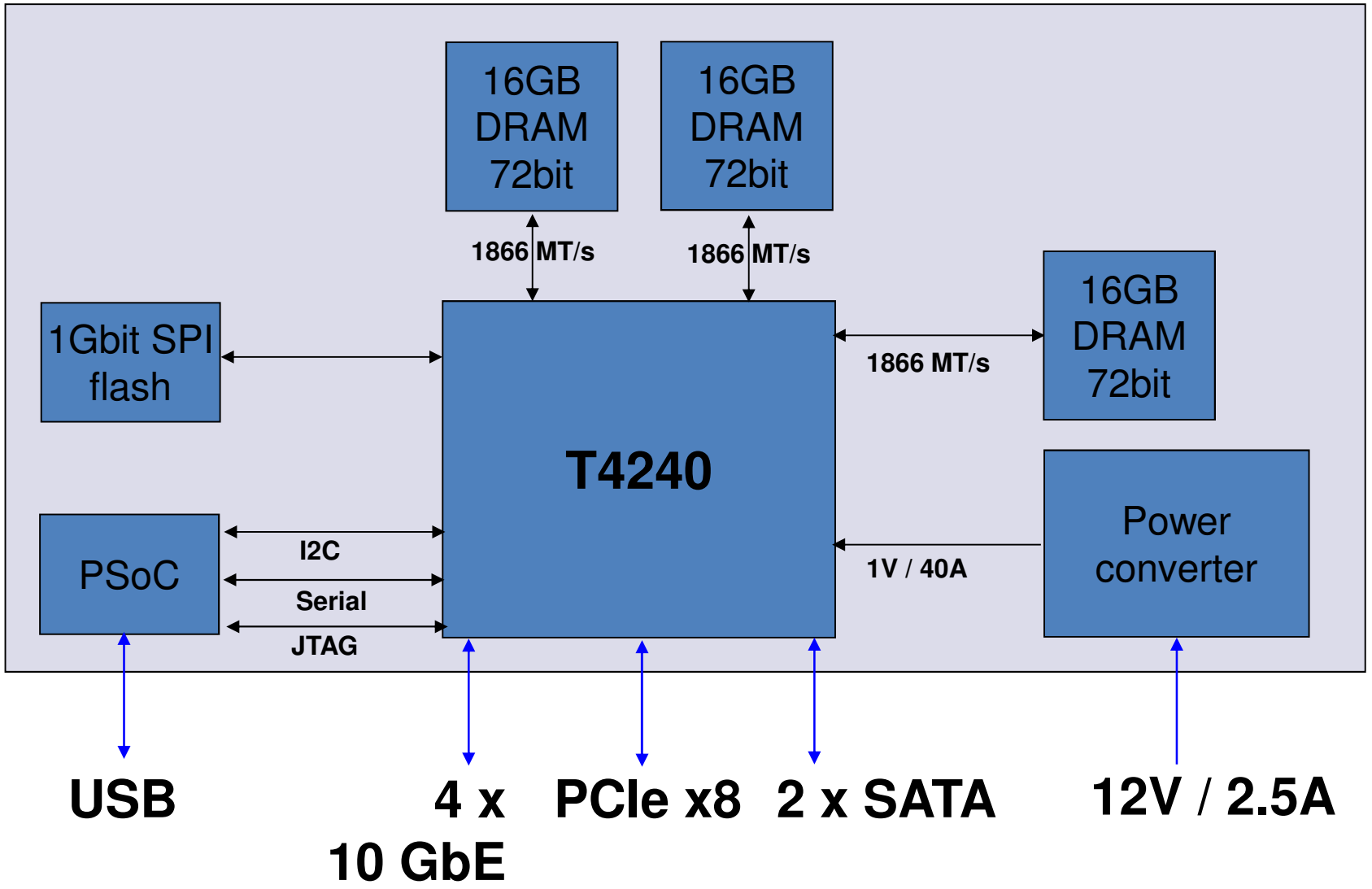
Built for *Embedded* market

Impressive power management features

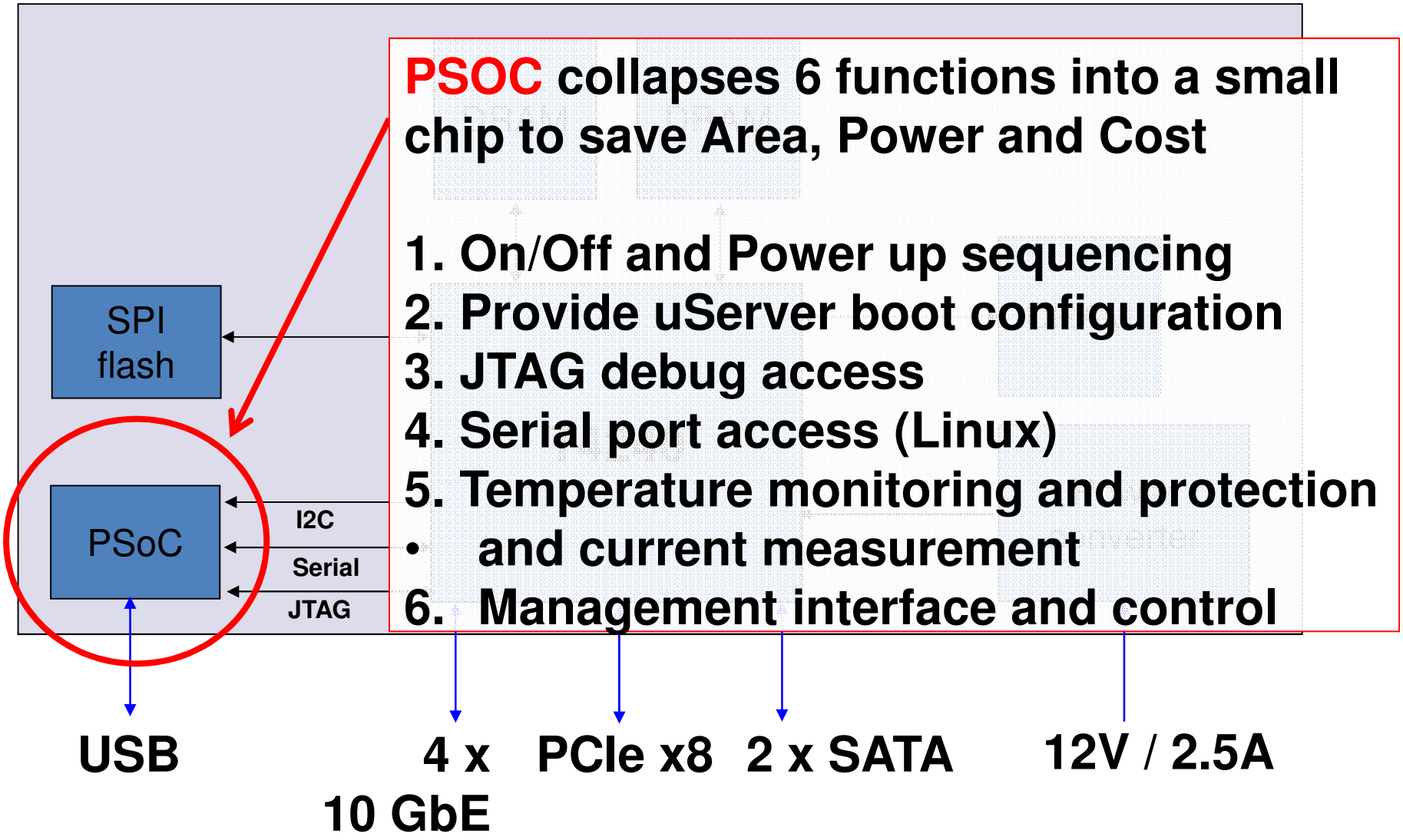
Not great for HPC:
not enough DP-FP units
No DDR prefetching



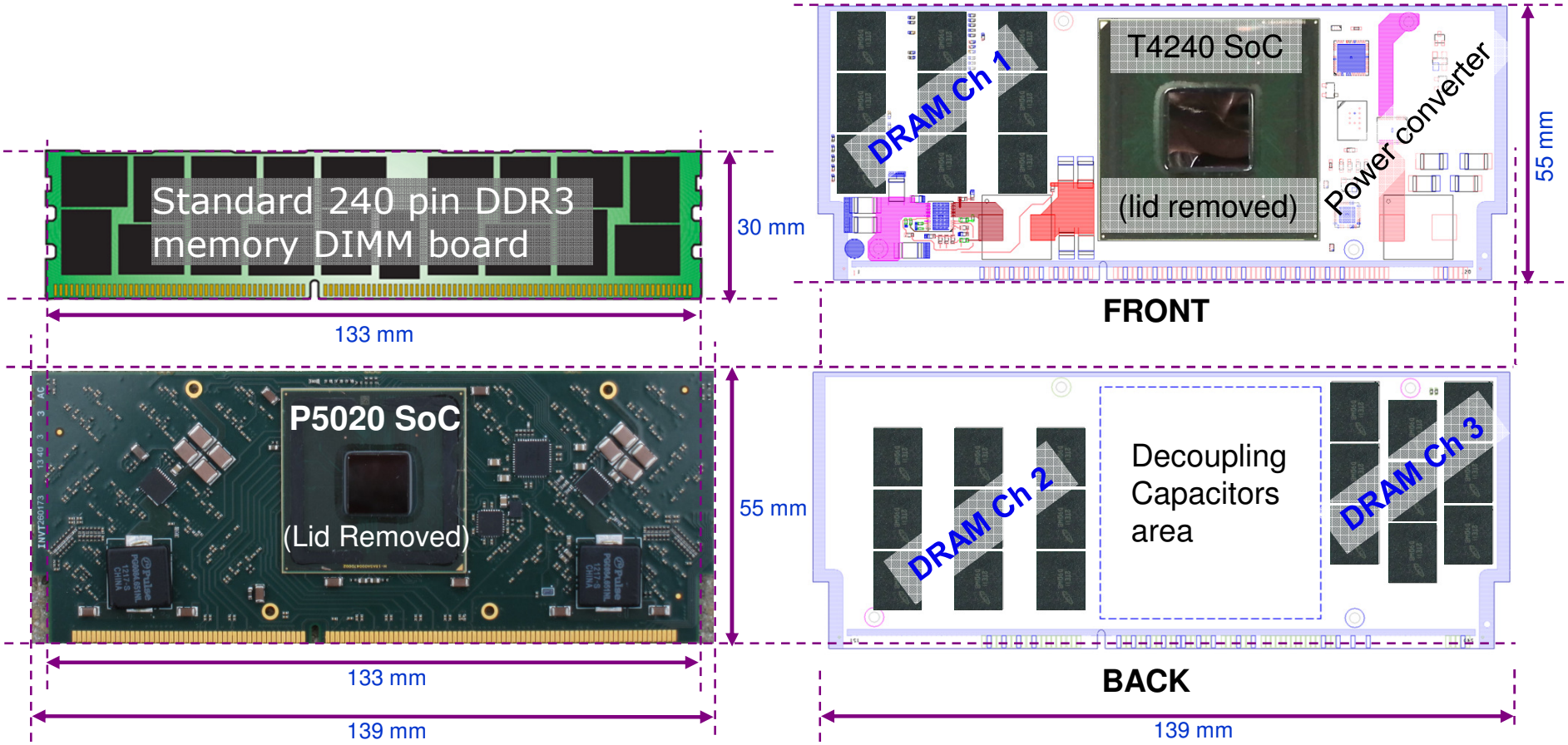
DOME compute node board diagram



DOME compute node board diagram



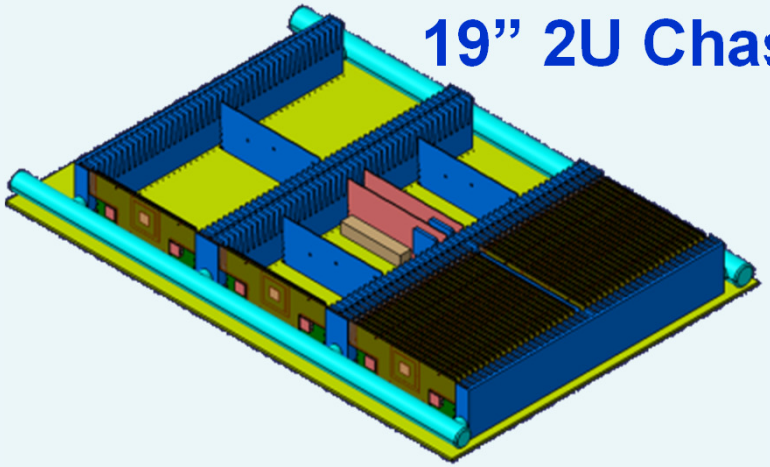
DOME Compute node board form factor



P5020/P5040 (Generation 1)

T4240 Generation 2

Planned System: 2U rack unit



19” 2U Chassis w/ Combined Cooling & Power

128 compute node boards

1536 cores / 3072 Threads

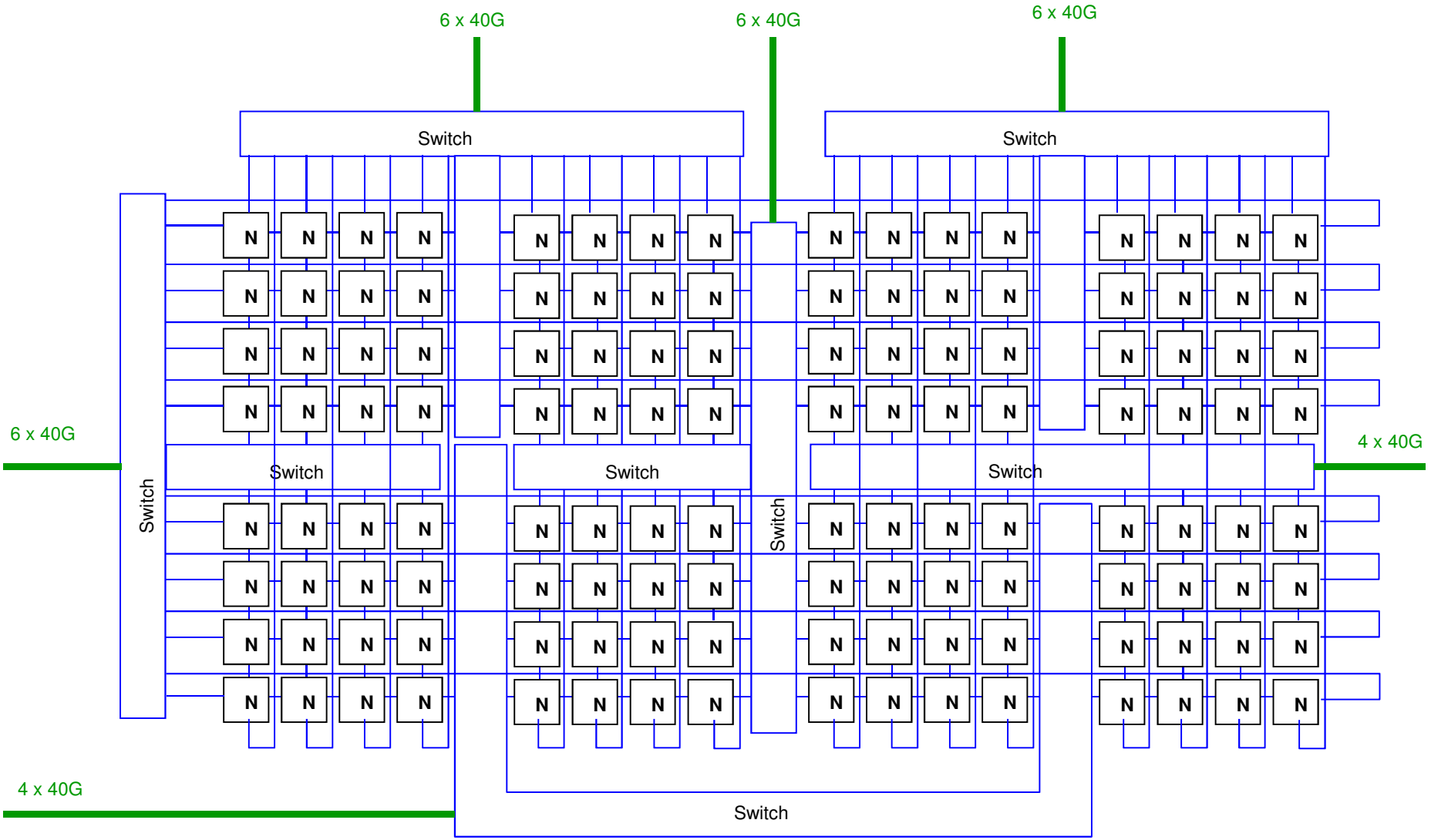
6 TB DRAM

1.28Tbps Ethernet (@40Gbps)

→ Datacenter-in-a-box

- Expected 2U unit total power: ~ 6kW
- Integrated mains power converter to 12V distribution: 12V / 500A
- Each compute node has own 12V / 40W converter
- Common Power Converter boards for all other supplies
- High radix 10GbE / 40GbE switch boards (under construction)
- Connects to Mains, Rack level Water, 32x 40Gbps Ethernet
- **Hot-water cooled for efficiency and density**

Planned network for 128 nodes with 40G external links



- **32 external 40G ports using Ethernet switches**
- **1280 Gbps external BW**

Performance Measurement Results

CPU	Freescale **T4240** 12 cores; 24 thr. 28nm Bulk	Intel Xeon **E3-1230L v3** 4 cores; 8 threads 22nm FinFet
CPU2006 Benchmark Test Environment	System: T4240RDB-PB **1.666 GHz core clock,** **1.866 GT/s 6GB DRAM, 3 channels** Fedora 20, Kernel 3.12.19 GCC 4.7.2 gcc options: -O3 -mcpu=powerpc64	System: Supermicro X10SAE **1.8 GHz core clock; Turbo disabled** **1.666 GT/s 8 GB DRAM, 2 channels** Fedora 19, Kernel 3.13.9 GCC 4.8.2 gcc options: -O3 -march=native -mtune=native
CINT-base – 1 thread	6.86	20.7
CINT-base – all threads	**109.34 (24 threads)**	**77.6 (8 threads)**
Coremark - all threads	**188K (24 threads)**	**65K (8 threads)**

Performance Measurement Results

CPU	Freescale **T4240** 12 cores; 24 thr. 28nm Bulk	Intel Xeon **E3-1230L v3** 4 cores; 8 threads 22nm FinFet
CPU2006 Benchmark Test Environment	System: T4240RDB-PB **1.666 GHz core clock,** **1.866 GT/s 6GB DRAM, 3 channels** Fedora 20, Kernel 3.12.19 GCC 4.7.2 gcc options: -O3 -mcpu=powerpc64	System: Supermicro X10SAE **1.8 GHz core clock; Turbo disabled** **1.666 GT/s 8 GB DRAM, 2 channels** Fedora 19, Kernel 3.13.9 GCC 4.8.2 gcc options: -O3 -march=native -mtune=native
CINT-base – 1 thread	6.86	20.7
CINT-base – all threads	**109.34 (24 threads)**	**77.6 (8 threads)**
Coremark - all threads	**188K (24 threads)**	**65K (8 threads)**

40% more performance @ 70% of node level energy consumption → **2x more operations per Watt**

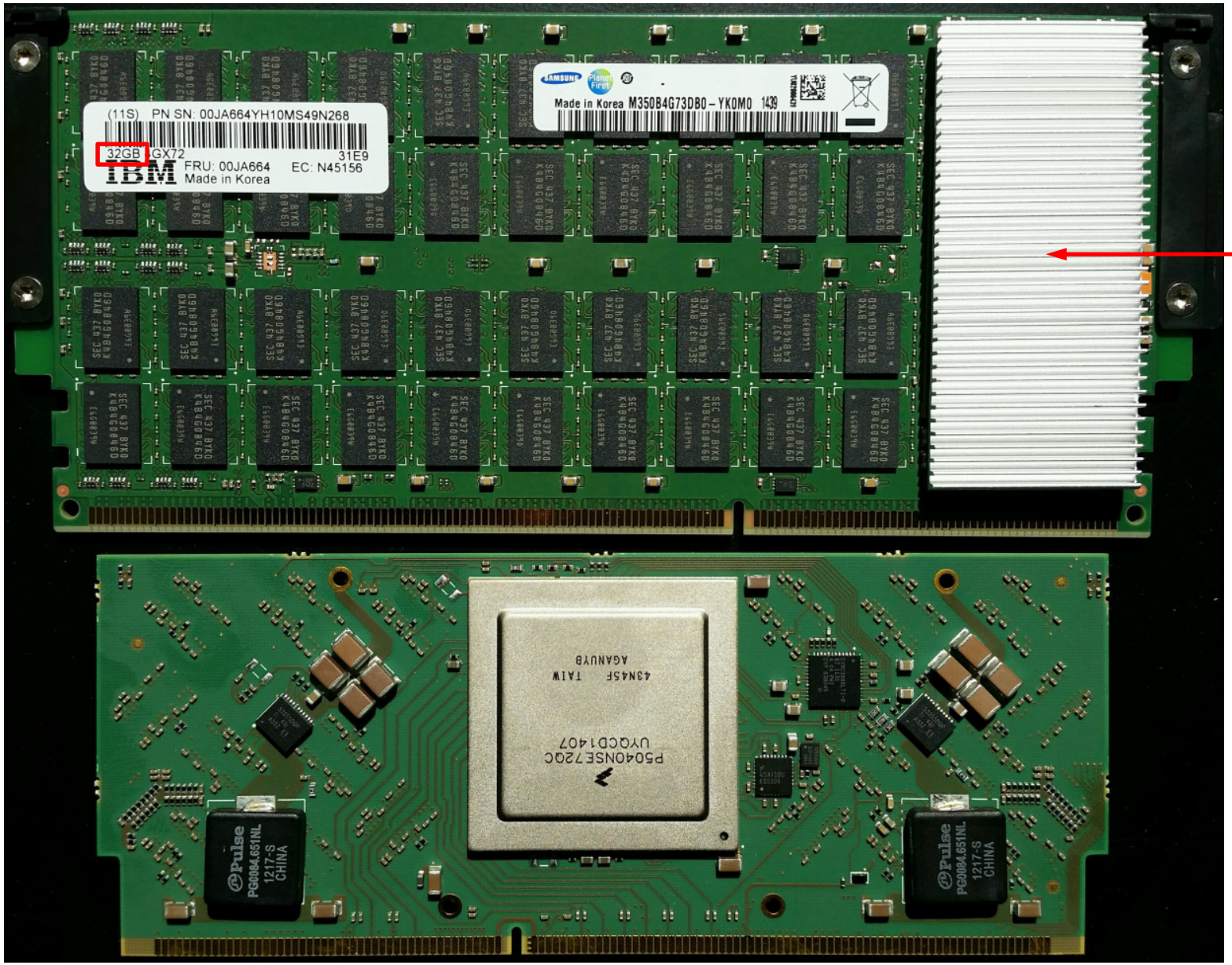
Performance Measurement Results

CPU	Freescale **T4240** 12 cores; 24 thr. 28nm Bulk	Intel Xeon **E3-1230L v3** 4 cores; 8 threads 22nm FinFet
CPU2006 Benchmark Test Environment	System: T4240RDB-PB **1.666 GHz core clock,** **1.866 GT/s 6GB DRAM**, 3 channel Fedora 20, Kernel 3.12.18 GCC 4.7.2 gcc options: -O3 -m32 -mcpu=powerpc64	System: Supermicro X10SAE **3.3 GHz core clock; Turbo disabled** **1.6 GT/s 8 GB DRAM**, 2 channel Fedora 19, Kernel 3.13.9 GCC 4.8.2 gcc options: -O3 -march=native -mtune=native
CINT-base – 1 thread	6.86	20.7
CINT-base – all threads	**171 (24 threads)**	**77.6 (8 threads)**
Coremark – all threads	**188K (24 threads)**	**65K (8 threads)**

Innovators Dilemma at work
The incumbents don't get it

40% more performance @ 70% of node level energy consumption → **2x more operations per Watt**

Comparison



P8 memory DIMM

?

DOME compute node

Power Measurement Results

Power measurement on rev 1 board #5, on 7 + 8 april 2015; PSoC firmware 2-mar-15

current measurements at 12V input of power converters, T4240 temp < 65C

voltage domain current measured @ 12V input	**1V8 I/O**		**DRAM**		**1V0 core**		**total node**
condition	**mA**	**W**	**mA**	**W**	**A**	**W**	**W**
PSOC only power	3.4	0.0408	74	0.888	0.0008	0.0096	0.9384
T4240 power on, kept in reset	75	0.9	152	1.824	0.32	3.84	6.564
u-boot prompt (idle)	77.6	0.9312	350	4.2	1.48	17.76	22.8912
Linux prompt, idle system	77.6	0.9312	315	3.78	1.58	18.96	23.6712
BW_MEM 512M, 24 thr	77.3	0.9276	450	5.4	1.65	19.8	26.1276
stream, 24 thread	77.3	0.9276	470	5.64	1.65	19.8	26.3676
BW_MEM 512, 24 thr	77.7	0.9324	320	3.84	2.53	30.36	35.1324
idle at XCFE desktop	77.7	0.9324	320	3.84	1.6	19.2	23.9724
SpecInt PerlBench, 24 thr	77.8	0.9336	400	4.8	2.63	31.56	**37.2936**
SpecInt PerlBench, 12 thr	78	0.936	355	4.26	2.2	26.4	31.596
SpecInt gcc, 12 thr	78	0.936	416	4.992	1.7	20.4	26.328

Remarks

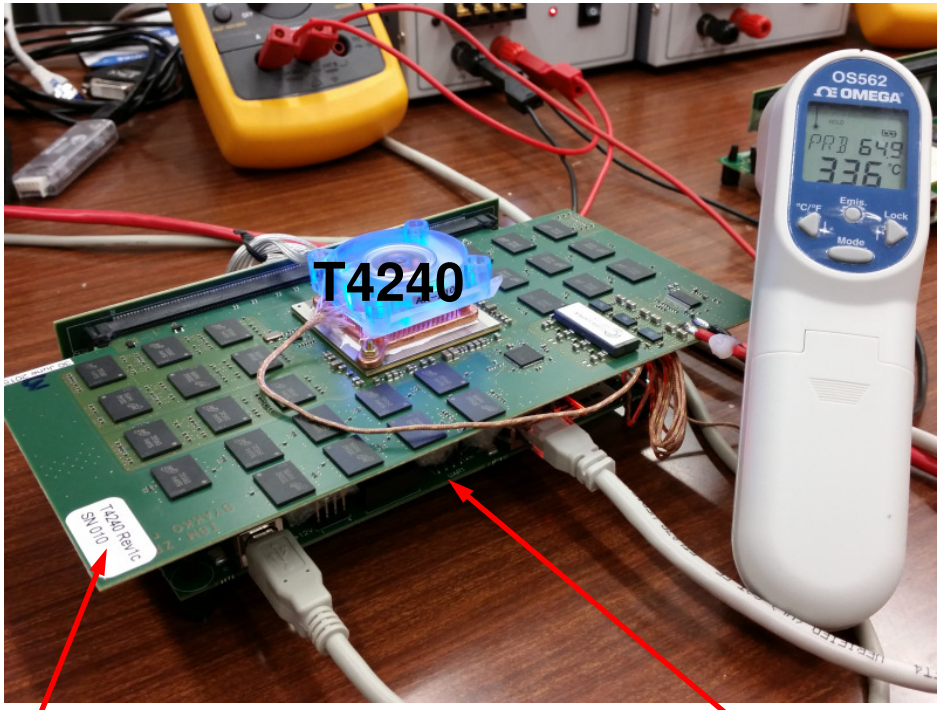
New **Big-Data** Metric: Memory BW density
→use raw memory BW available at SoC or CPU
→divide by volume of entire enclosure, incl. HDD, PCI slots

DOME 128node 2U rack unit: **159GB/s/Liter** (peak)
P8 server S822L (dual socket): **13.9GB/s/Liter** (peak)

- New era – perfect storm and Innovators Dilemma
- µServer is all about SoC and packaging

- This is a serendipitous data point

LIVE DEMO



We demonstrate a single node running:

- Fedora 20
- XFCE Desktop
- Stream
- CPMD

- And… live 1V domain current measurement

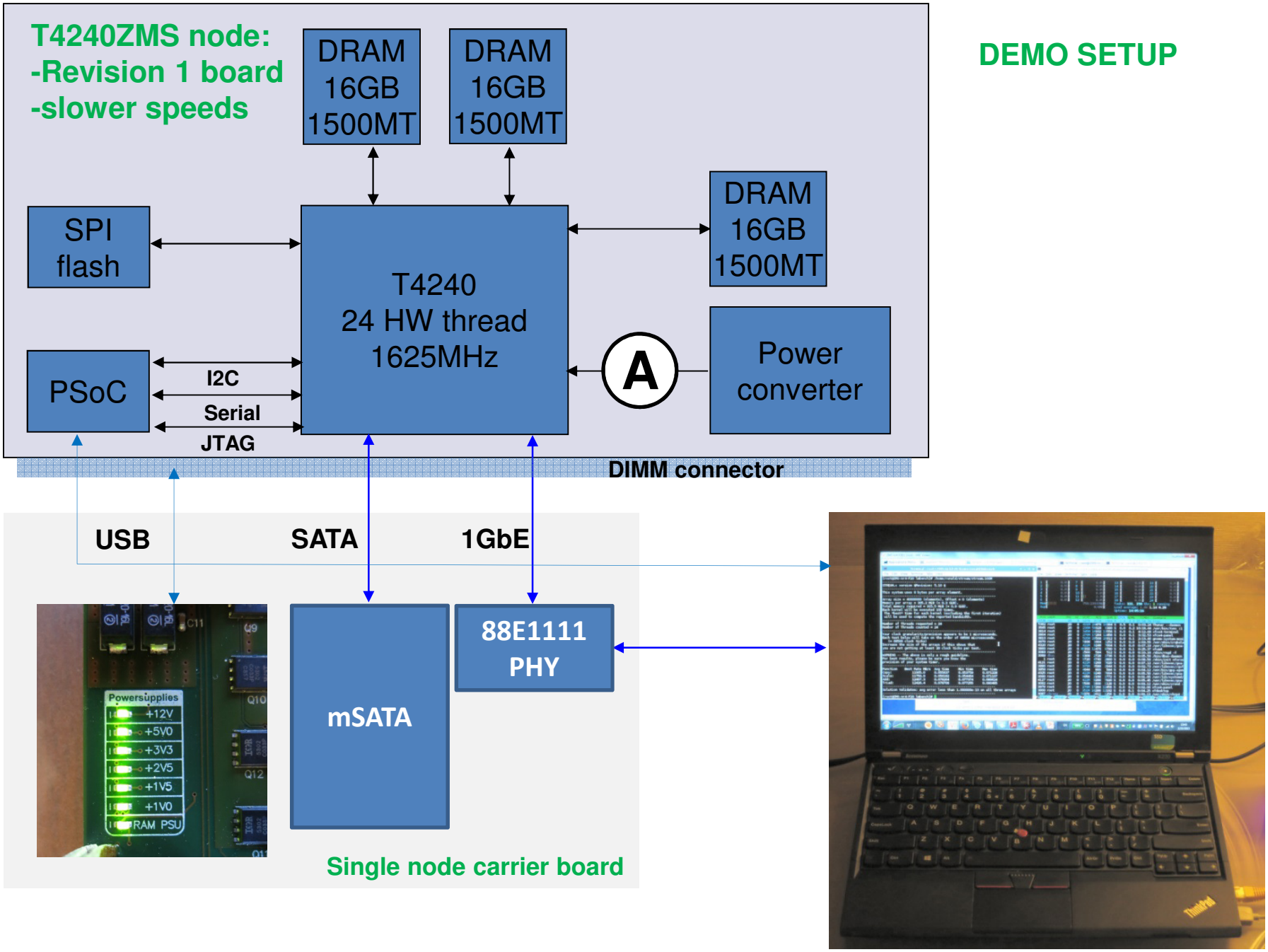
compute node

mini BaseBoard

Showing a revision-1 board T4240ZMS compute server:

- Larger than DOME form factor, same netlist
- All components on top side (save bring-up time and expense)
- Air-cooled for single node operation

DEMO SETUP


T4240ZMS node:
-Revision 1 board
-slower speeds
DRAM 16GB 1500MT
DRAM 16GB 1500MT
DRAM 16GB 1500MT
SPI flash
T4240
24 HW thread
1625MHz
A
Power converter
PSoC
I2C
Serial
JTAG
DIMM connector
USB
SATA
1GbE
mSATA
88E1111 PHY
Powersupplies
+12V
+5V0
+3V3
+2V5
+1V5
+1V0
RAM PSU
Single node carrier board

Status and Plans

Until YE 2015

2016 a new compute node

Beyond 2016

H2020 proposals

•1. Please provide a brief overview of the activities at your Institution that address the technical challenges in hardware and software architecture. These efforts can be in traditional scientific HPC, or in the area of "Big Data" and Data Analytics. You have an opportunity to highlight unique perspectives you can bring to the workshop as representatives of the broader International community.
•Analytics, HPC (alg; codes; arch), Accelerators, Security

2. A key goal of the BDEC workshop is to systematically map the opportunities for Big Data synergy with Extreme-Scale HPC. In recent decades, the HPC community has used HPC systems that were created from the integration of commodity computing components that were largely designed and developed for the much larger desktop and server markets. Moving forward, in an analogous manner it is very likely that future HPC systems will be created from the integration of commodity computing components that were originally designed for the much larger Big Data markets. Do you agree with this statement and from your perspective are there other synergies that can be leveraged?
•uServer is using embedded market commodity SoC – example of other leveraged synergy

3. What are your priorities for international cooperation in designing and developing hardware and software architectures for both Big Data and Extreme-scale Computing? From the perspective of your Institution, do you have examples of successful cooperation or collaboration? Examples can be cited as workshops you hosted, successful open source technology collaborations, visiting researcher positions, joint papers, etc.
•Successful collaboration with ASTRON, influencing SKA. Great collab with FSL. **DOME USER PLATFORM**
Have developed low cost – high performance data aggregator to feed IoT data into HPC – opportunity here!

4. In what areas could you benefit from contributions provided by other institutions including industry vendors, academia, and government organizations? Whether open source or proprietary, what would you seek in the way of hardware and software components and tools, experimental results and findings, or driver computational challenges from the world-wide HPC and big data community to further your own goals in these emerging cooperative fields?
•The insights in this project (it's the system design, stupid) tell us what SoC our community should build…
Looking for 100M$ to build better SoC – I have ideas….

Links

SKA: http://www.skatelescope.org
DOME: http://www.dome-exascale.nl
µServer: http://www.zurich.ibm.com/microserver
T4240 system: http://swissdutch.ch:6999
Wikipedia: https://en.wikipedia.org/wiki/Microserver
Twitter: https://twitter.com/ronaldgadget

Videos:

Impossible µServer: http://t.co/4vEkEVEazO
Innovators Dilemma: http://youtu.be/imweQe8NgnI
DOME T4240 Fedora: http://youtu.be/D6da5DqcyQk

Literature

“Energy-Efficient Microserver Based on a 12-Core 1.8GHz 188K-CoreMark 28nm Bulk CMOS 64b SoC for Big-Data Applications with 159GB/s/L Memory Bandwidth System Density”, R.Luijten et al., ISSCC15, San Francisco, Feb 2015

“The DOME embedded 64 bit microserver demonstrator”, R. Luijten and A. Doering, ICICDT 2013, Pavia, Italy, May 2013

“**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.

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
- Alessandro Curioni, IBM ZRL
- Stephan Paredes, IBM ZRL
- Matteo Cossale, IBM ZRL
- James Nigel, FSL
- Boris Bialek, IBM Toronto
- Marco de Vos, Astron NL
- Vipin Patel, IBM Fishkill
- And many more remain unnamed….

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

Questions???

PS. I like lightweight things
µServer website: www.swissdutch.ch