- Click in the field ``c'down'' above to set a countdown.
- Click on the button below to start the countdown and the local time.

Start/Stop Timer

The QPACE 2 Project

People involved:

Paul Arts, Jacques Bloch, Hans Deinhart, Peter Georg, Benjamin Glaessle, Simon Heybrock, Yu Komatsubara, Robert Lohmayer, Simon Mages, Bernhard Mendl, Nils, Meyer Alessio Parcianello, Dirk Pleiter, Florian Rappl, Mauro Rossi, Norbert Sommer, Giampietro Tecchiolli, Tilo Wettig, Gianpaolo Zanier

The problem:

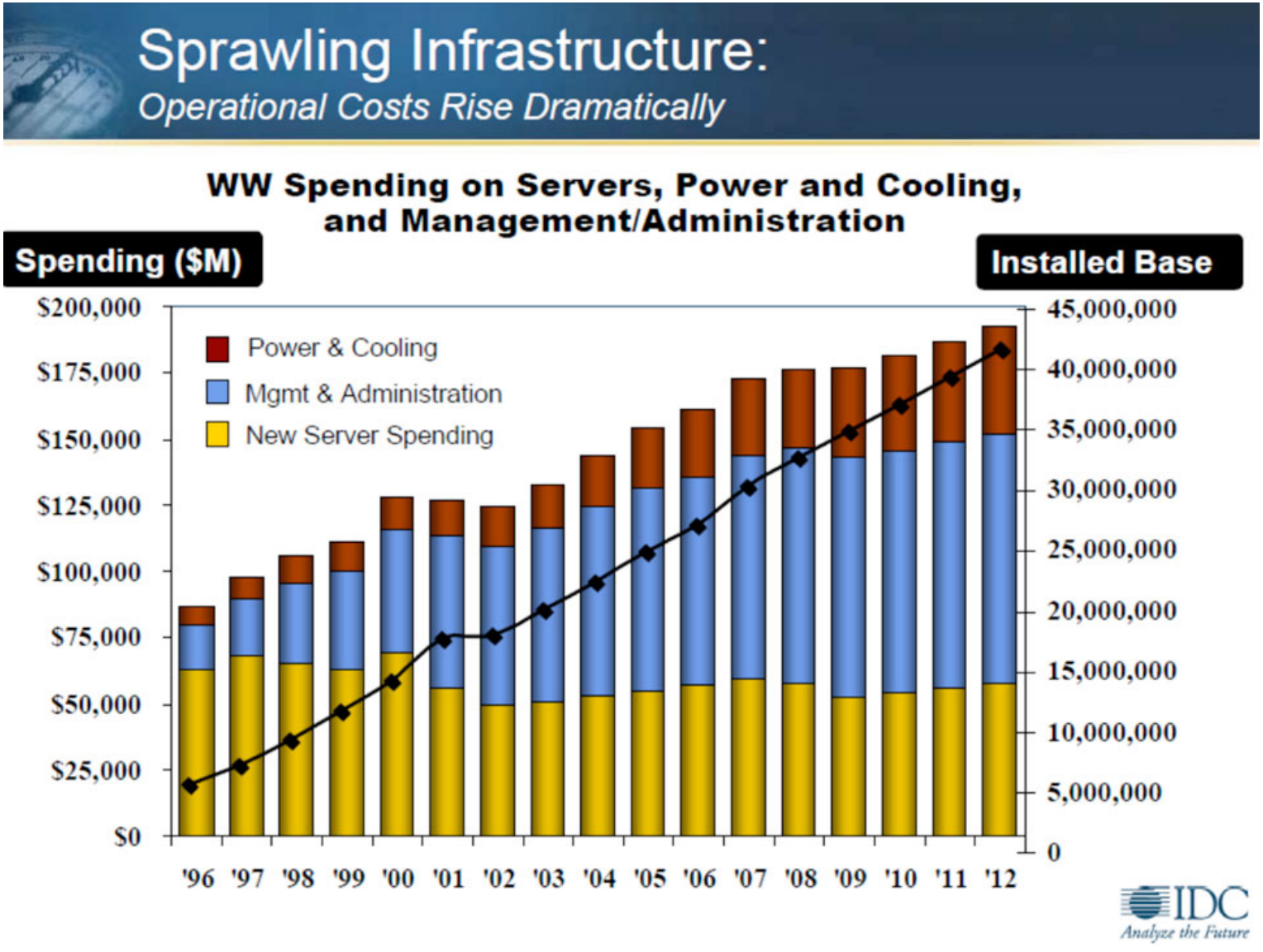
Costs for power & cooling have risen to more than half of the costs for new hardware:

year	new server	power & cooling
2011	48	26
1996	63	8

(Numbers in Billion USD, Source: IDC.com)

Cooling & energy **costs** become significant.

Cooling costs



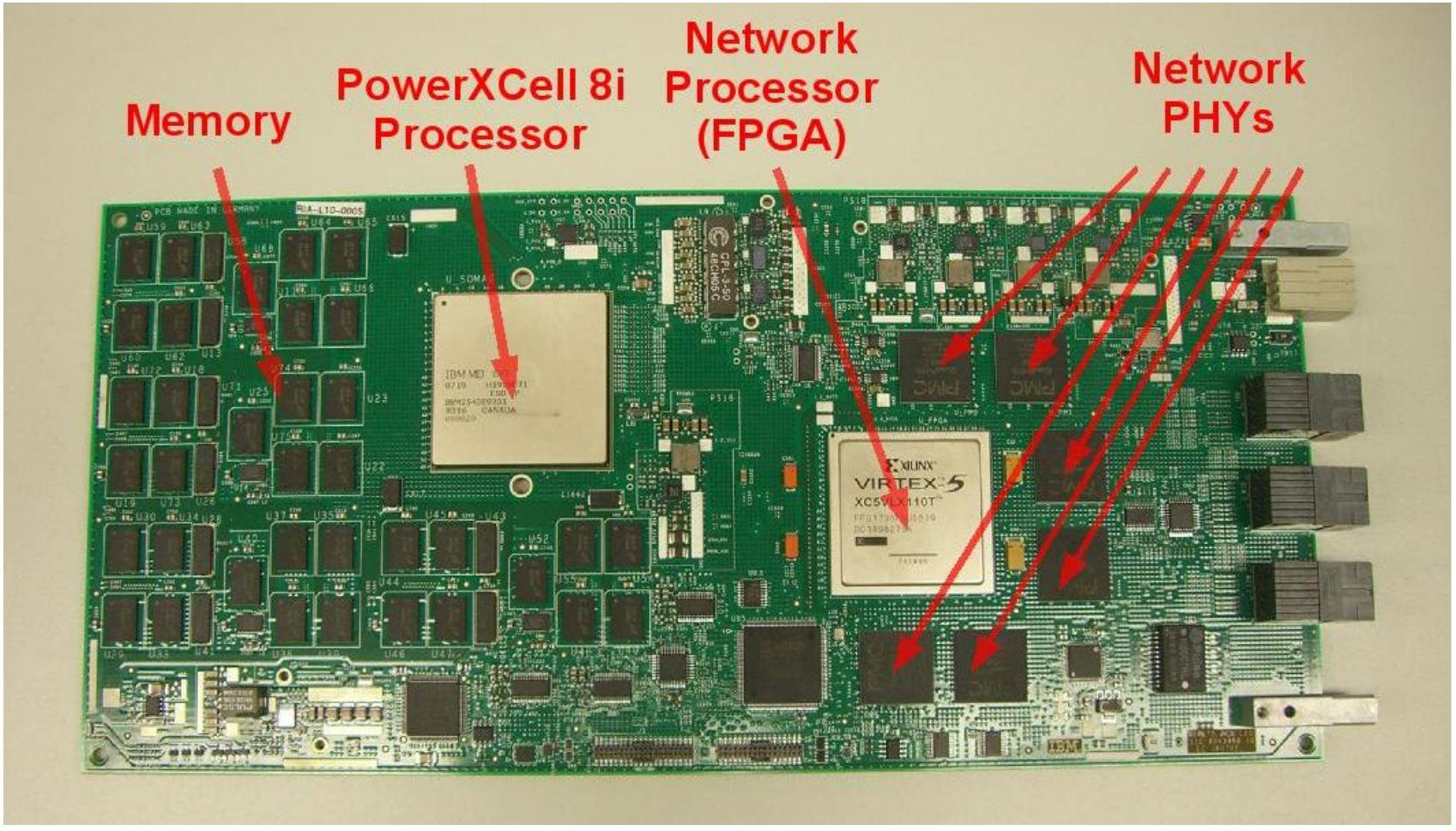
(Taken from crn.com)

Idea: Pair the most efficient processor with a network that matches our needs and efficient cooling.

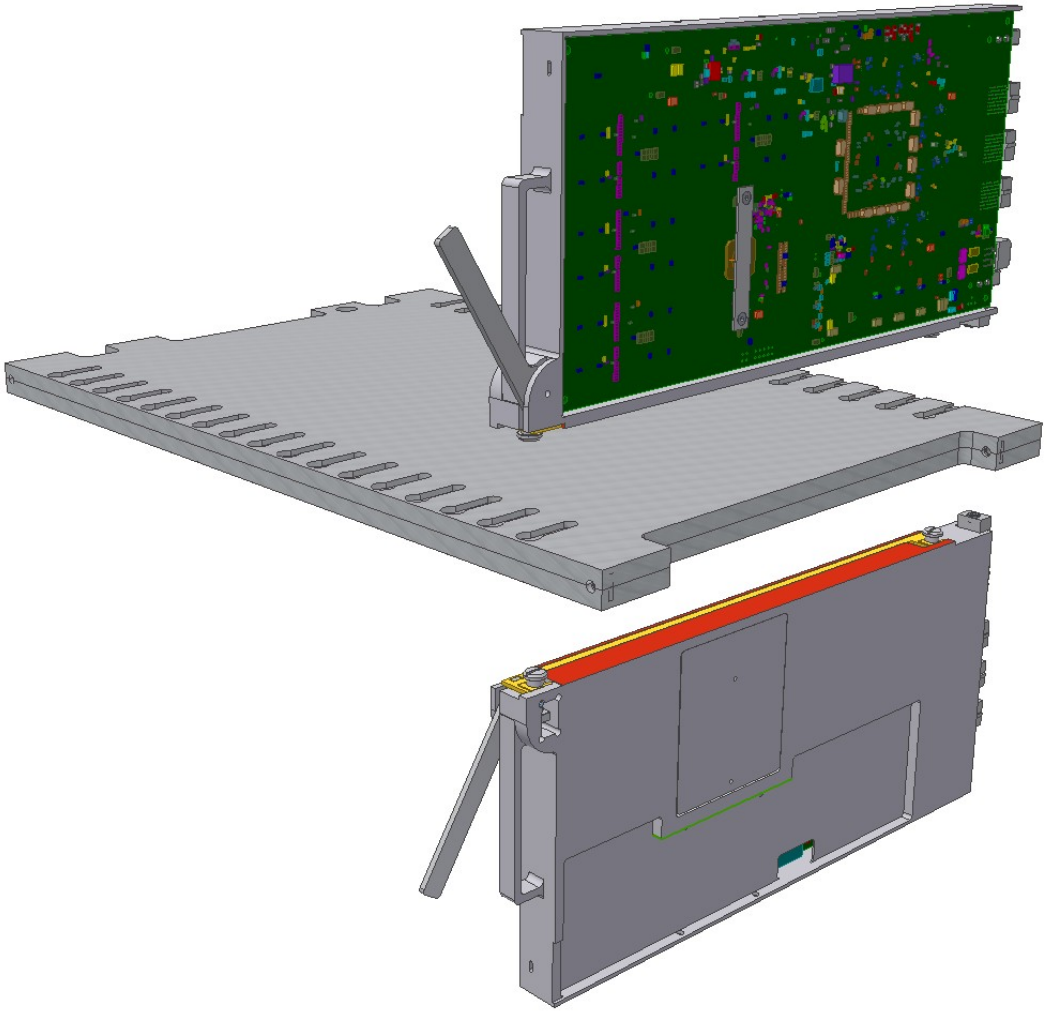
At that time, this meant:

- Cell Processor (IBM)
- Custom network
- indirect water cooling

Custom qpace node card



Indirect water cooling



Installation at Wuppertal and Jülich

Great success! Number 1 in the “Green500” list twice!

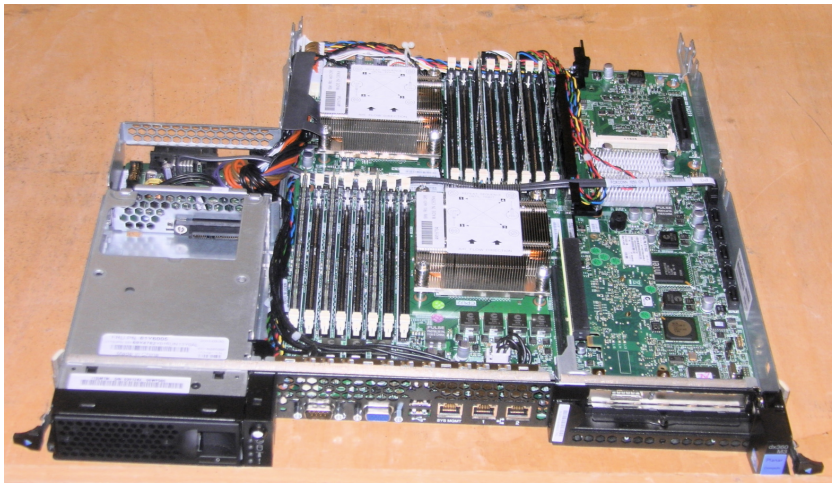
But this comes at a cost:

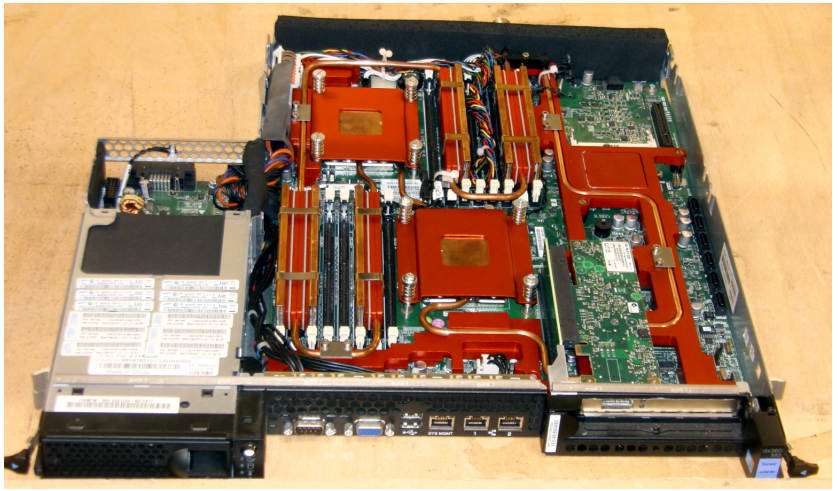
- Very tricky to program
- Development process involved a lot of manpower (but we did learn a lot)

Intermediate project with a different aim:

- Cold water cooling still costs money
- Need standard hardware,
- but aim for even higher energy efficiency: use hot water cooling

Joint Project with IBM: remodel standard hardware for hot water cooling (65°C inlet temperature)





Achievements:

- Energy reuse w/ adsorption chiller
- running stable with high water temperatures

QPACE 2 — Goals

Our gloas for QPACE 2:

- Build the “best” machine for LQCD
- Be energy efficient
- Keep costs moderate

Processor: Intel's Knights Corner (KNC)

- 16 GBytes memory
- clock speed of 1.238 GHz
- 512 bit wide registers
- not bootable
- runs Linux
- PCIe2 endpoint

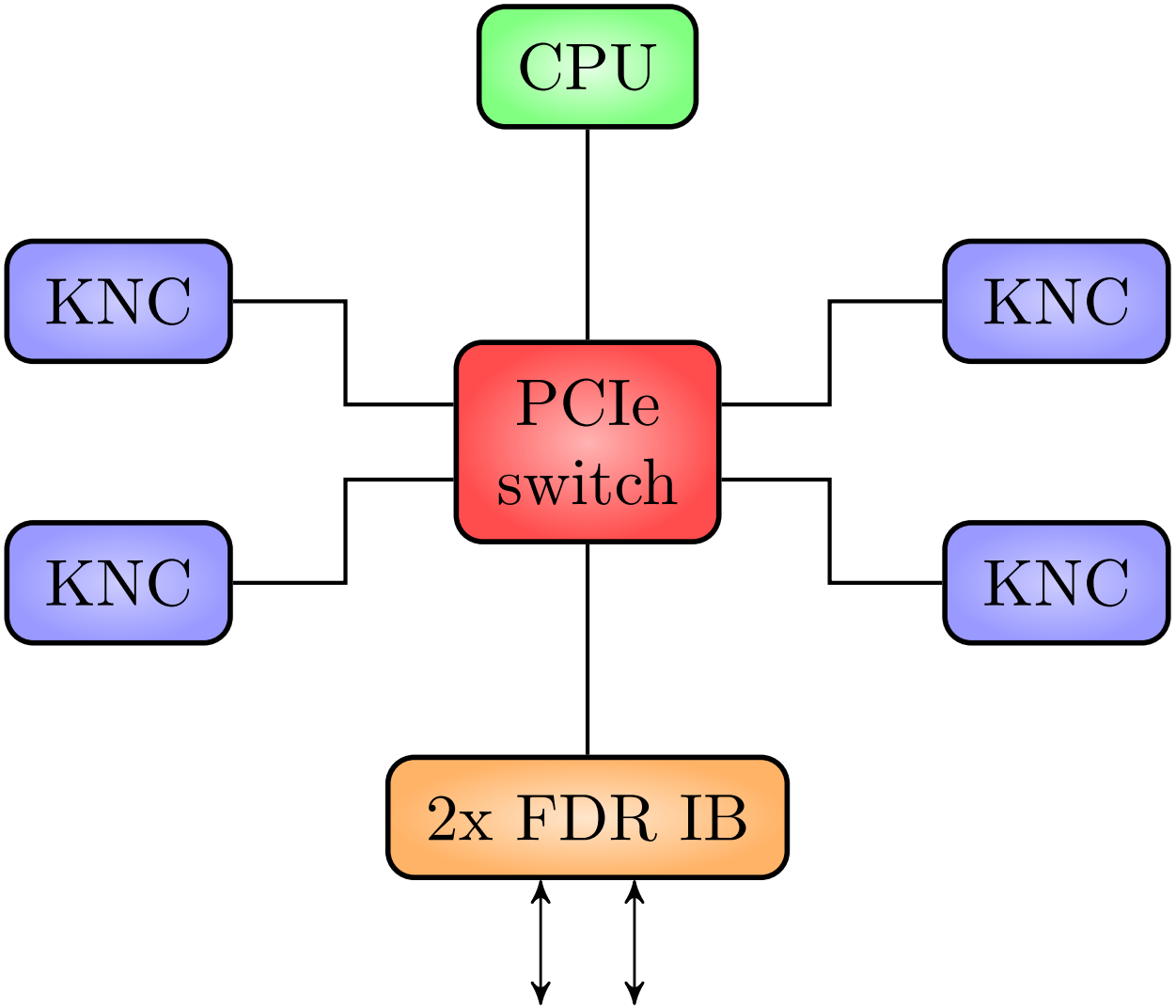
- 61 Cores
- peak performance of 1.2 TFlop/s (double precision)
- 4-way hyperthreading

Need to use at least 2 threads per core to get full performance, as instructions from a thread can only be issued every other cycle.

Main concept for a compute node:

- one efficient (weak) host CPU as PCIe root complex
- several KNCs
  4 in our case
- Infiniband HCA
  dual port FDR (2x 56 Gbit)
- connect everything with a PCIe switch

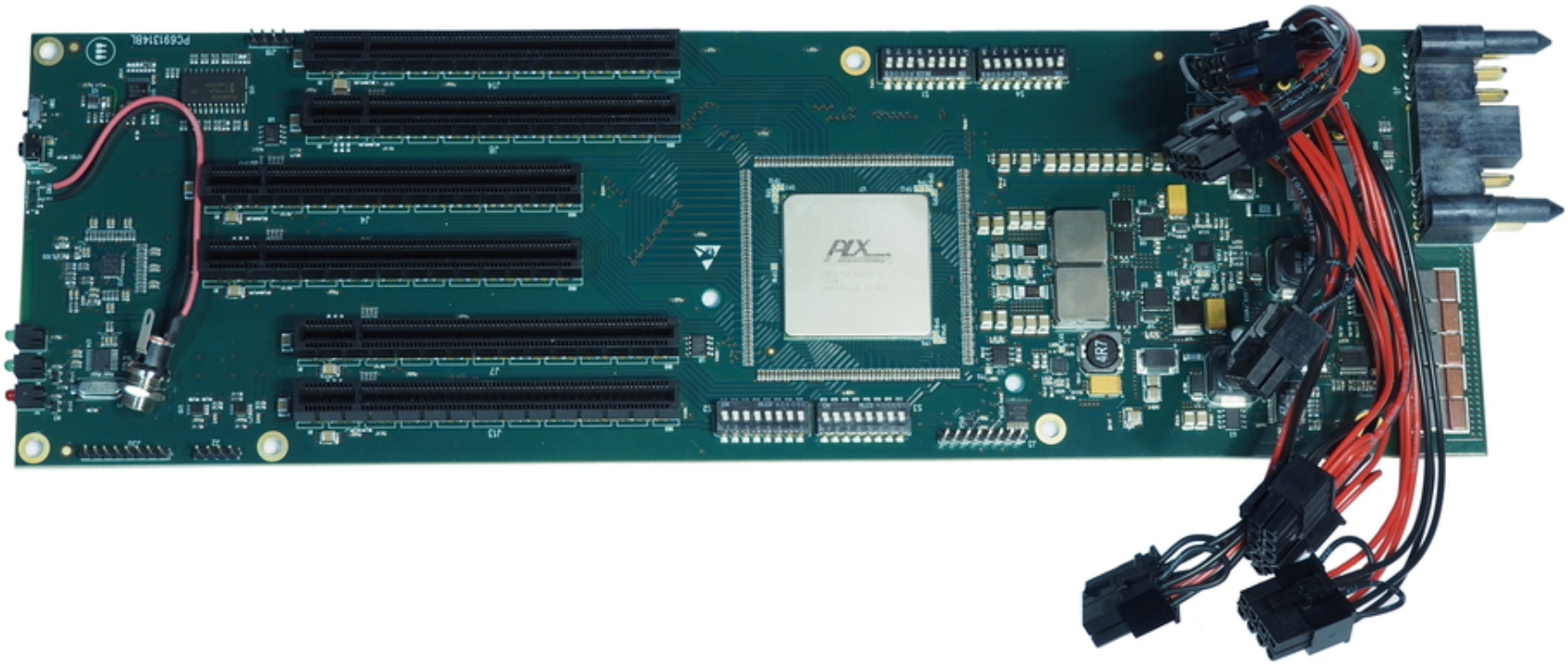
QPACE 2 — Node Scheme



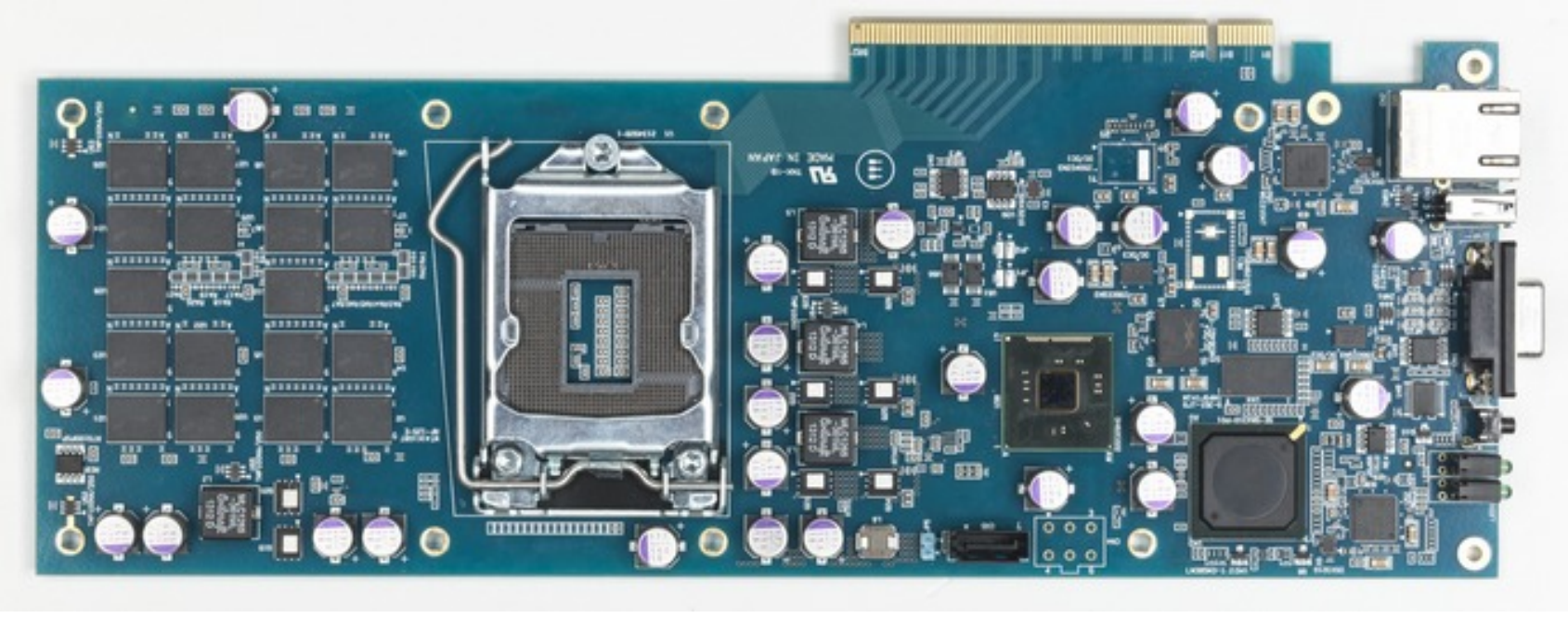
Board design, mechanical design:

- Industry partner: Eurotech (Amaro, Italy)
  - known from QPACE 1
  - board design
  - mechanical design
  - board manufacturing
- mechanical manufacturing: in-house

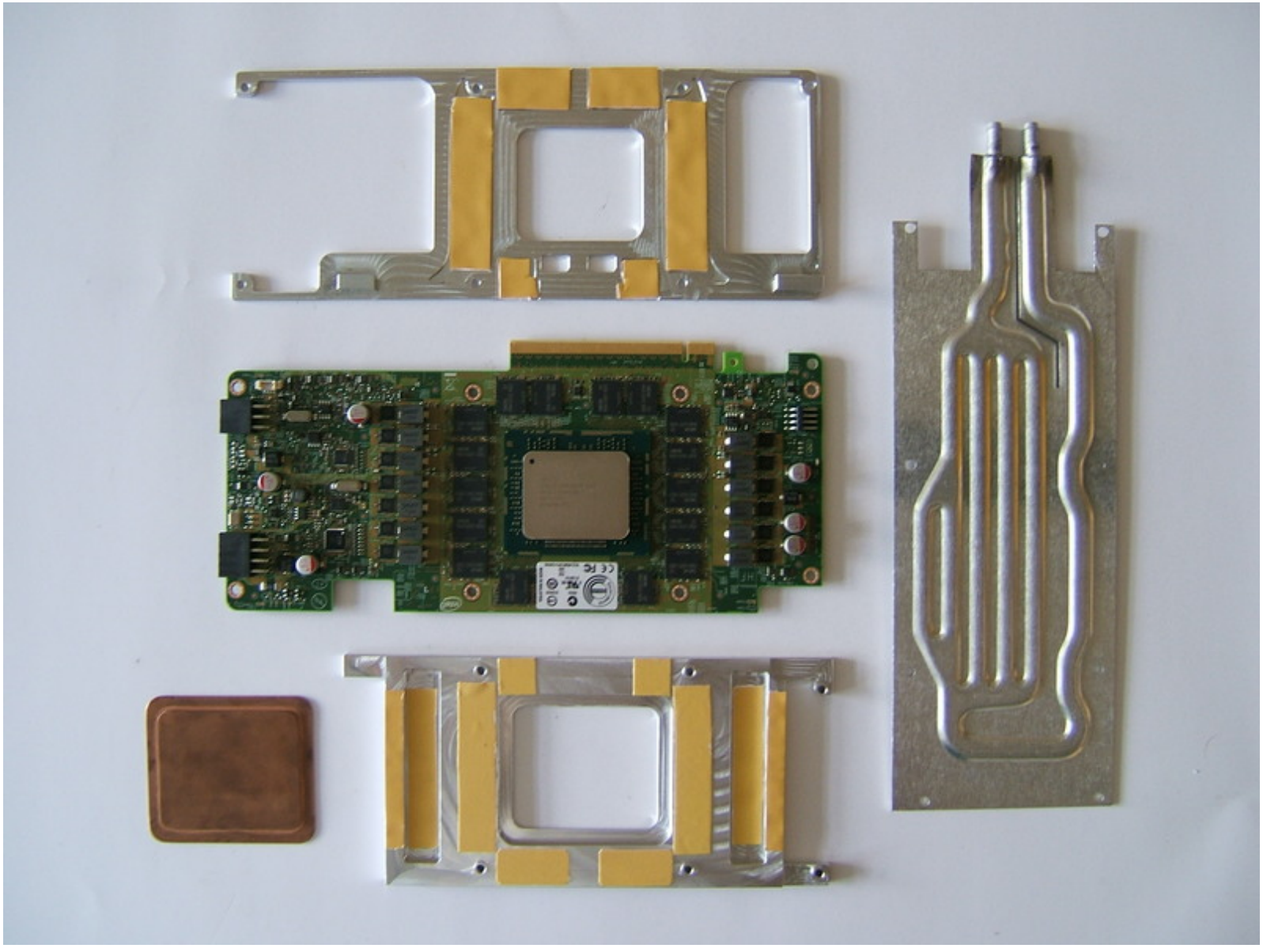
Midplane with PCIe switch and connectors



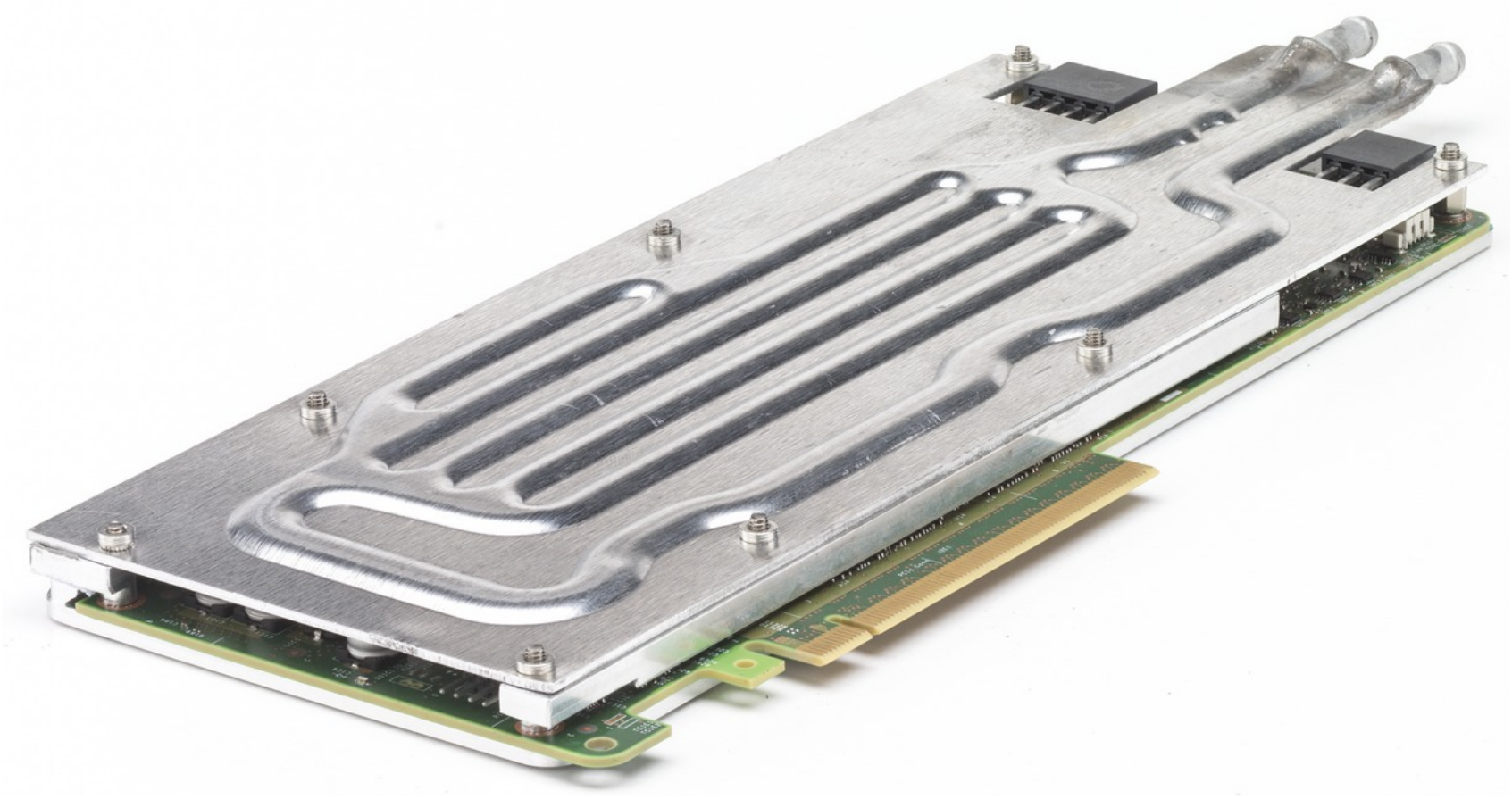
Host with CPU socket and BMC



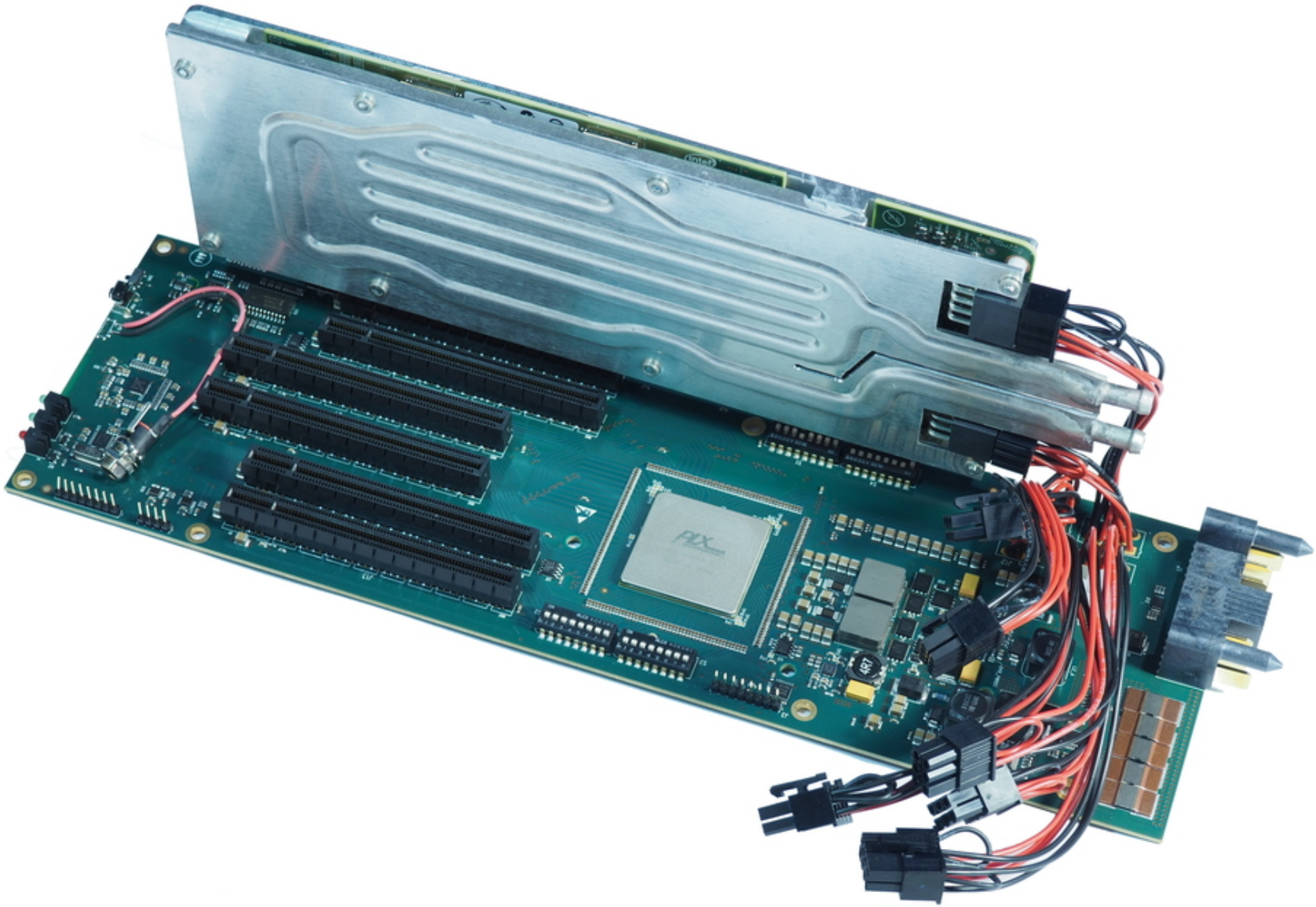
KNC, interposer and roll-bond heatsink



KNC sandwich (takes time & effort)



QPACE 2 — Node Pictures — KNC inserted

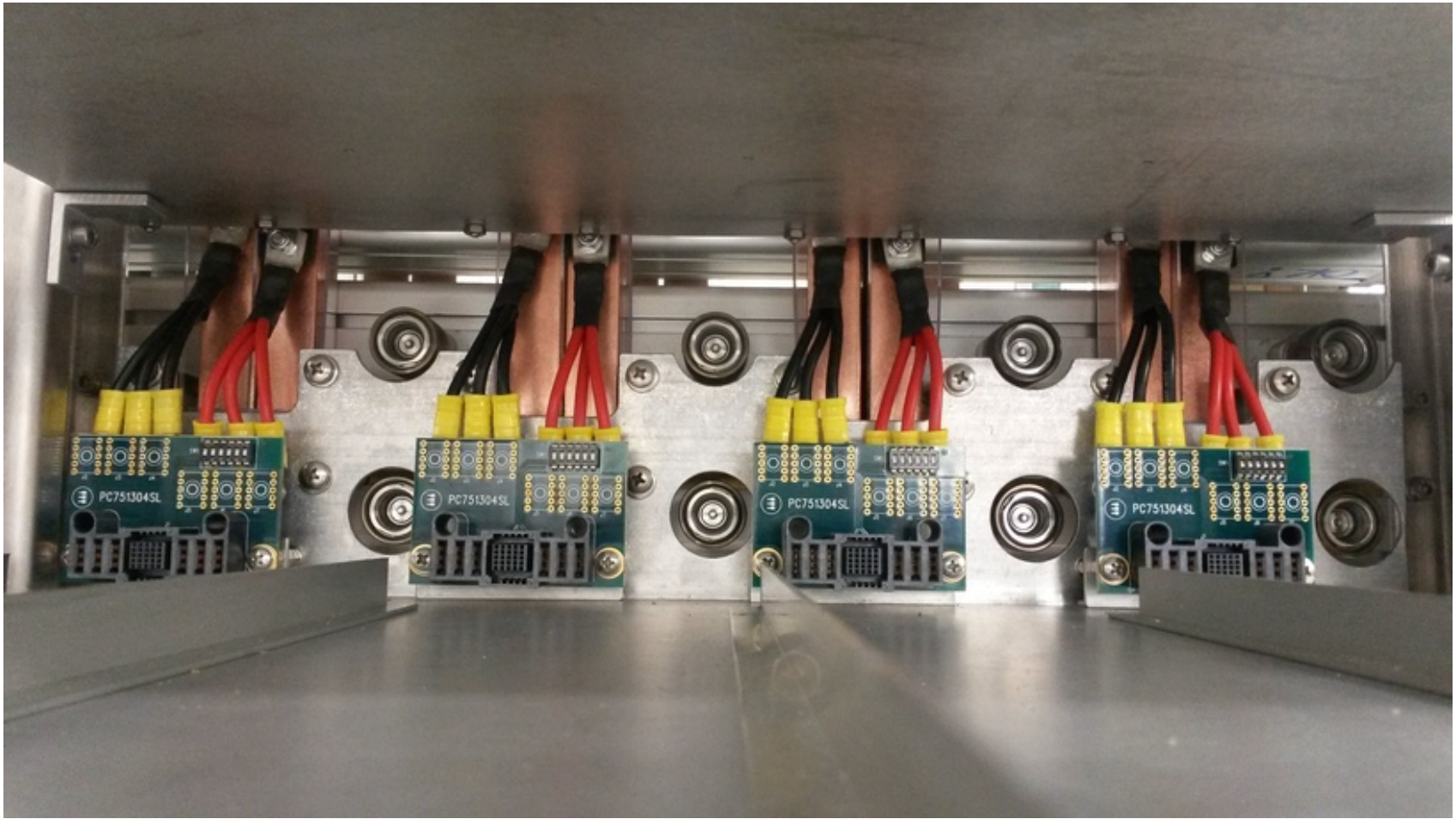


QPACE 2 — Node Pictures — Testing!

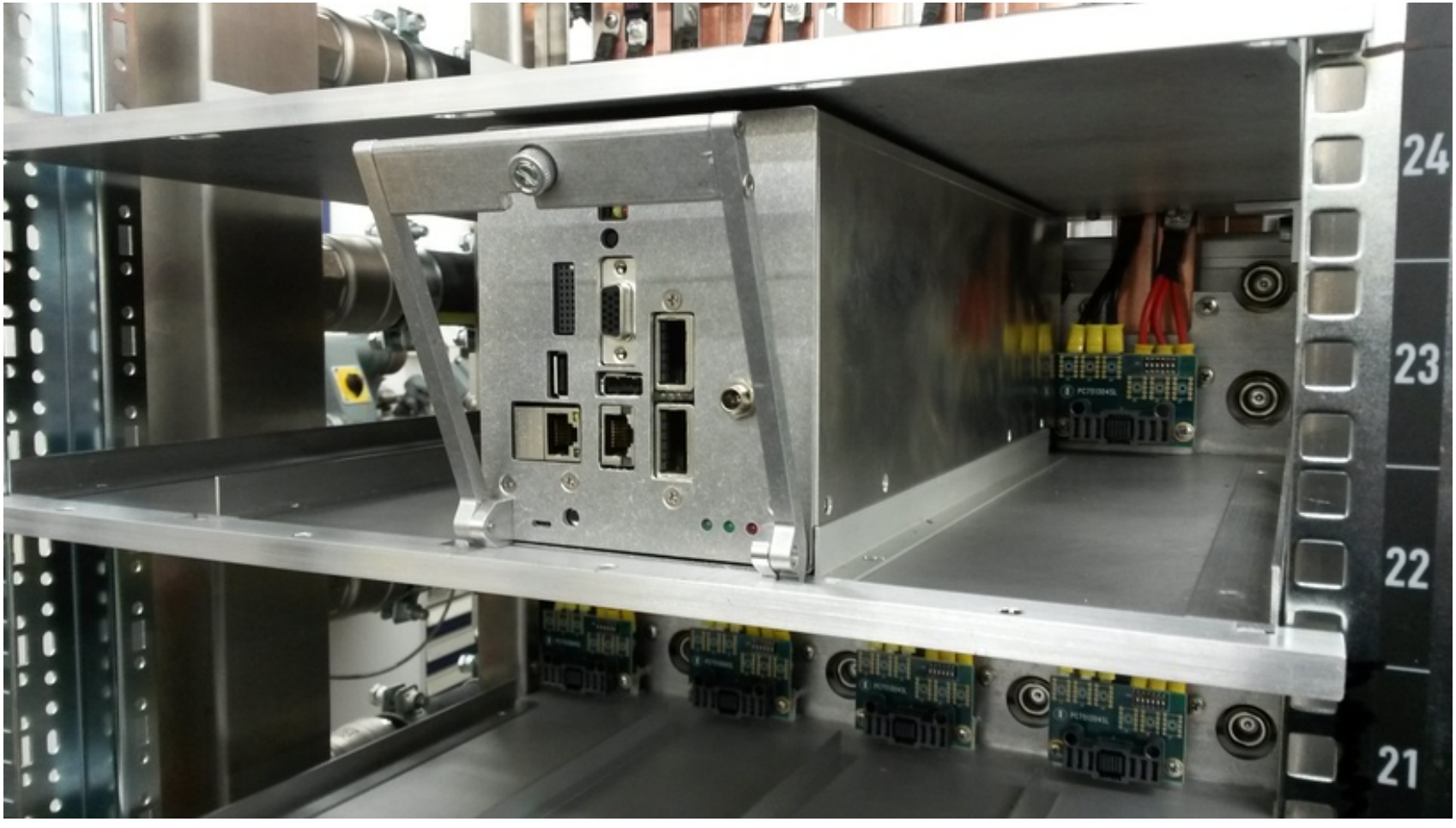
A “brick”

QPACE 2 — Node Pictures — housing (back)

QPACE 2 — Node Pictures — rack inside



QPACE 2 — Node Pictures — rack with node



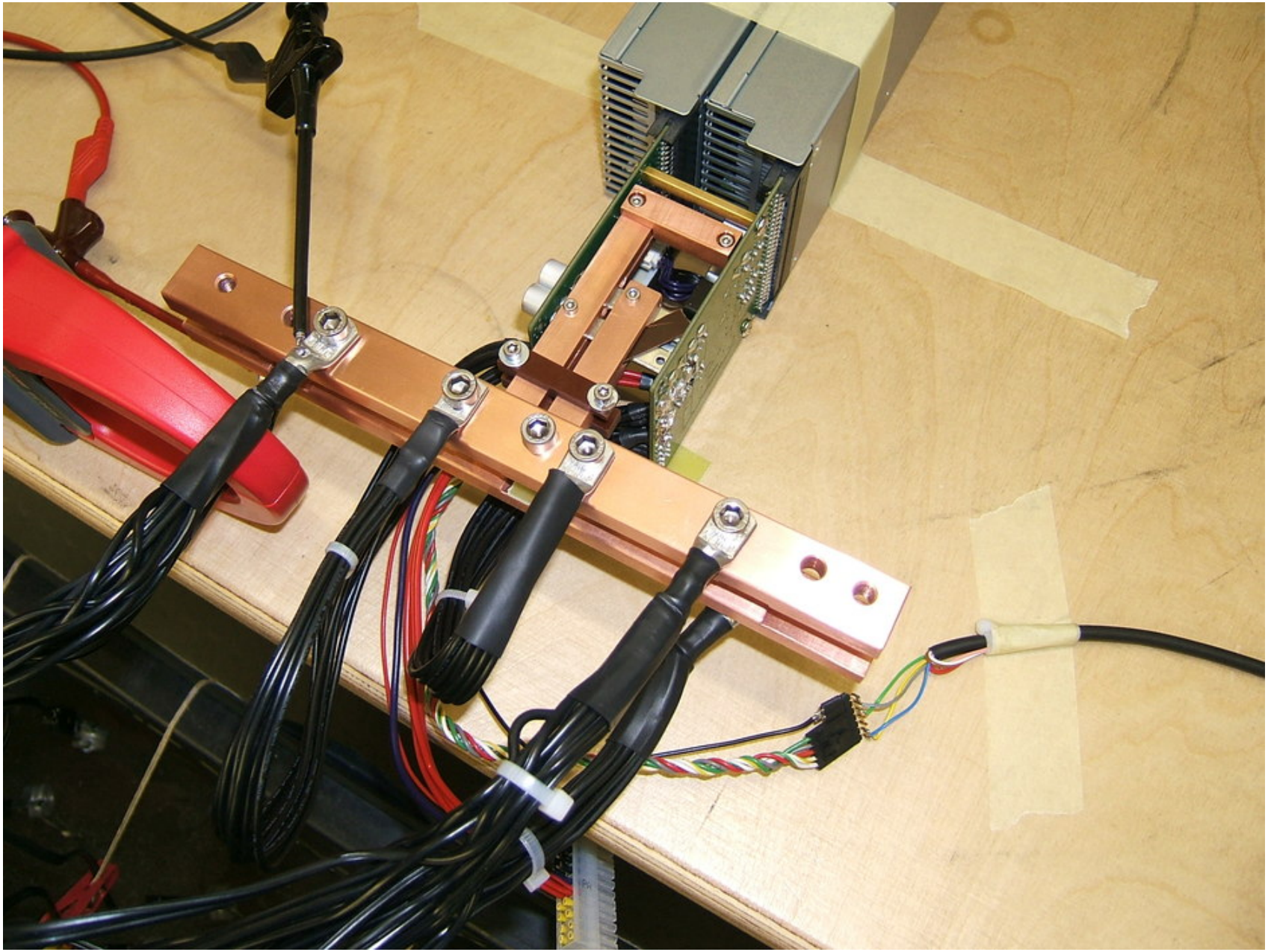
QPACE 2 — Node Pictures — power

QPACE 2 — whole system?

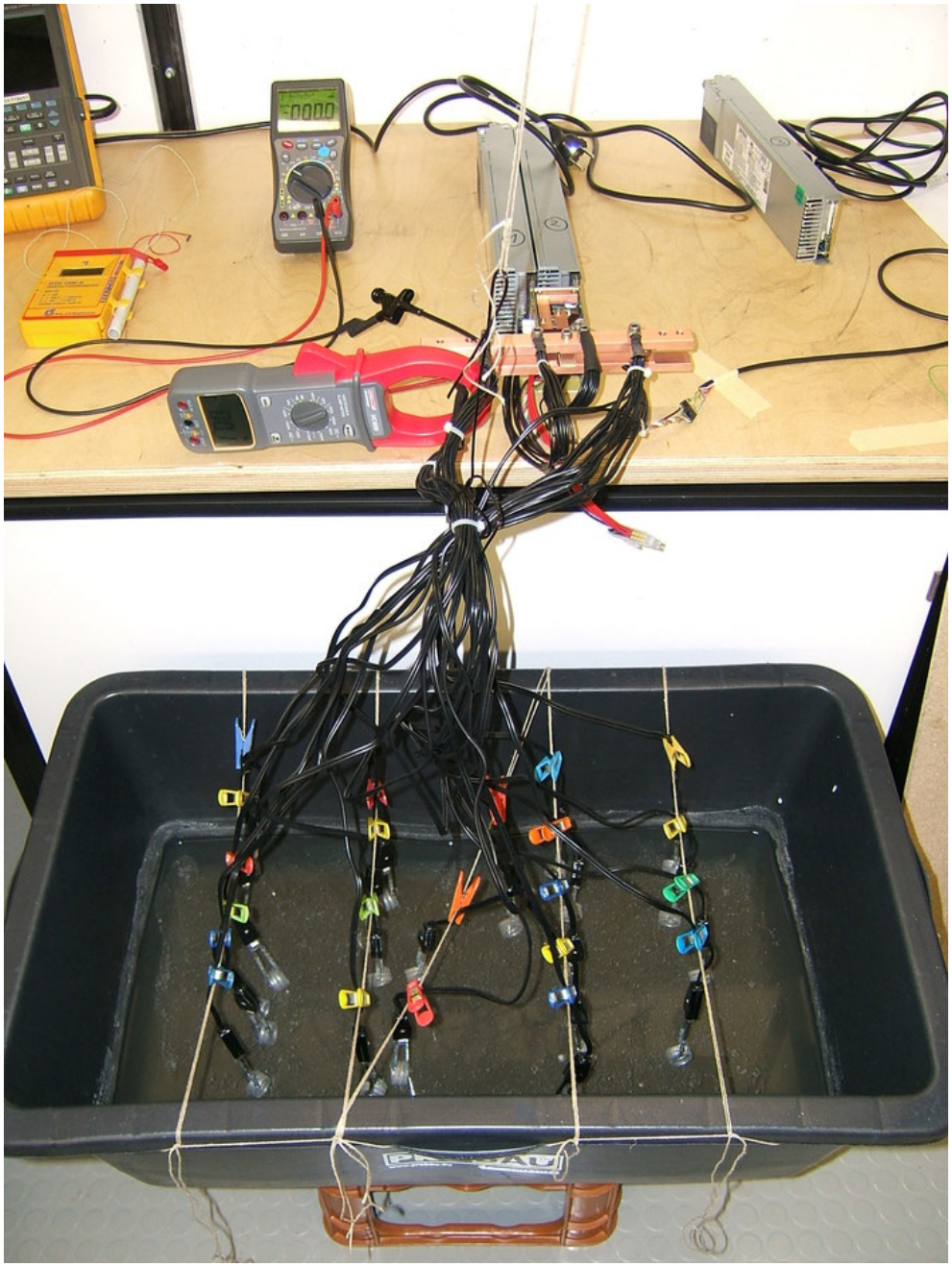
Put together *one* Rack

- 64 bricks
- makes 265 KNCs
- $\approx$ 300 TFLOP/S
- $\approx$ 80 kW power consumption

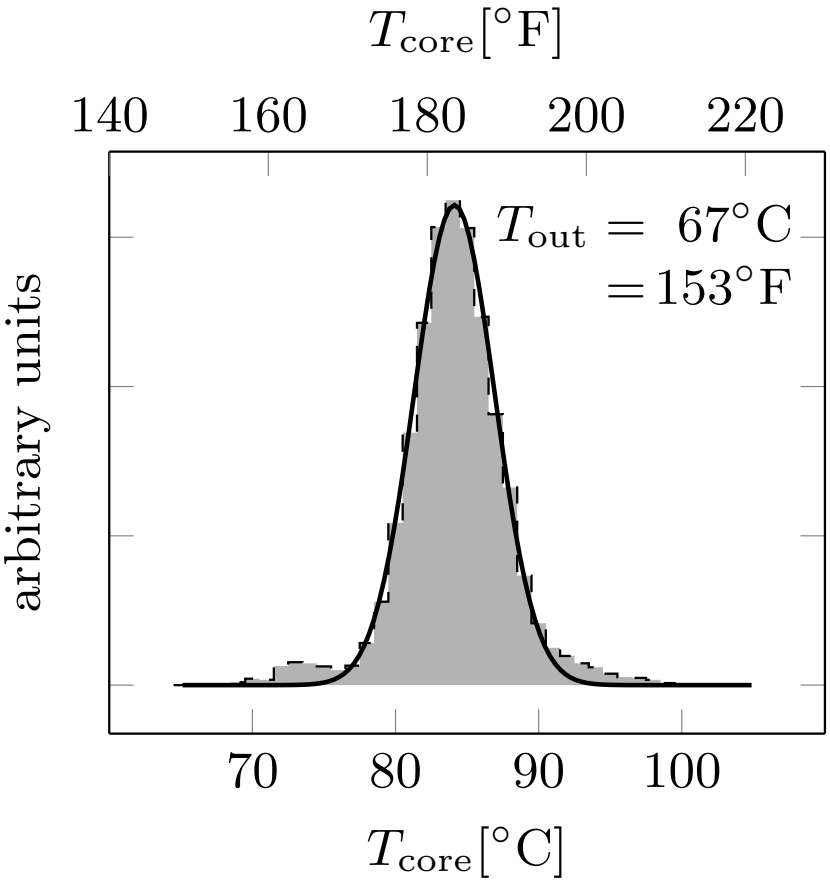
QPACE 2 — Node Pictures — power testing



QPACE 2 — Node Pictures — High performance camping



QPACE 2 — Lesson learned



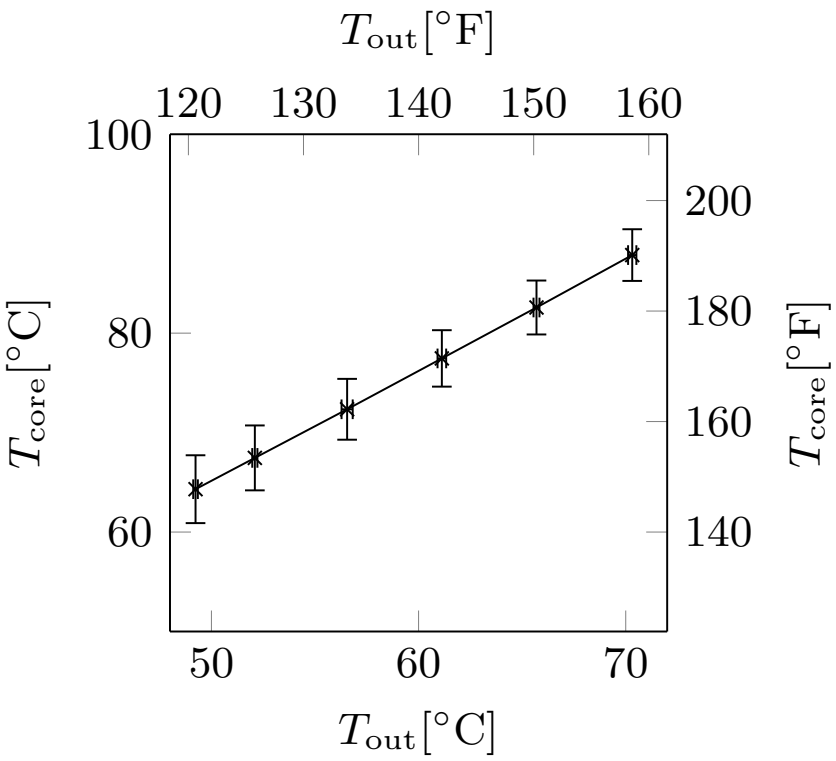
Lesson leared from idatacool: Temperature spread rather large.

Care for the “weak” nodes to avoid throttling.

Lecture Notes in Computer Science Volume 7905, 2013, pp 383-394 iDataCool: HPC with Hot-Water Cooling and Energy Reuse

arXiv: 1309.4887

QPACE 2 — Lesson learned



Lesson learned from idatacool:

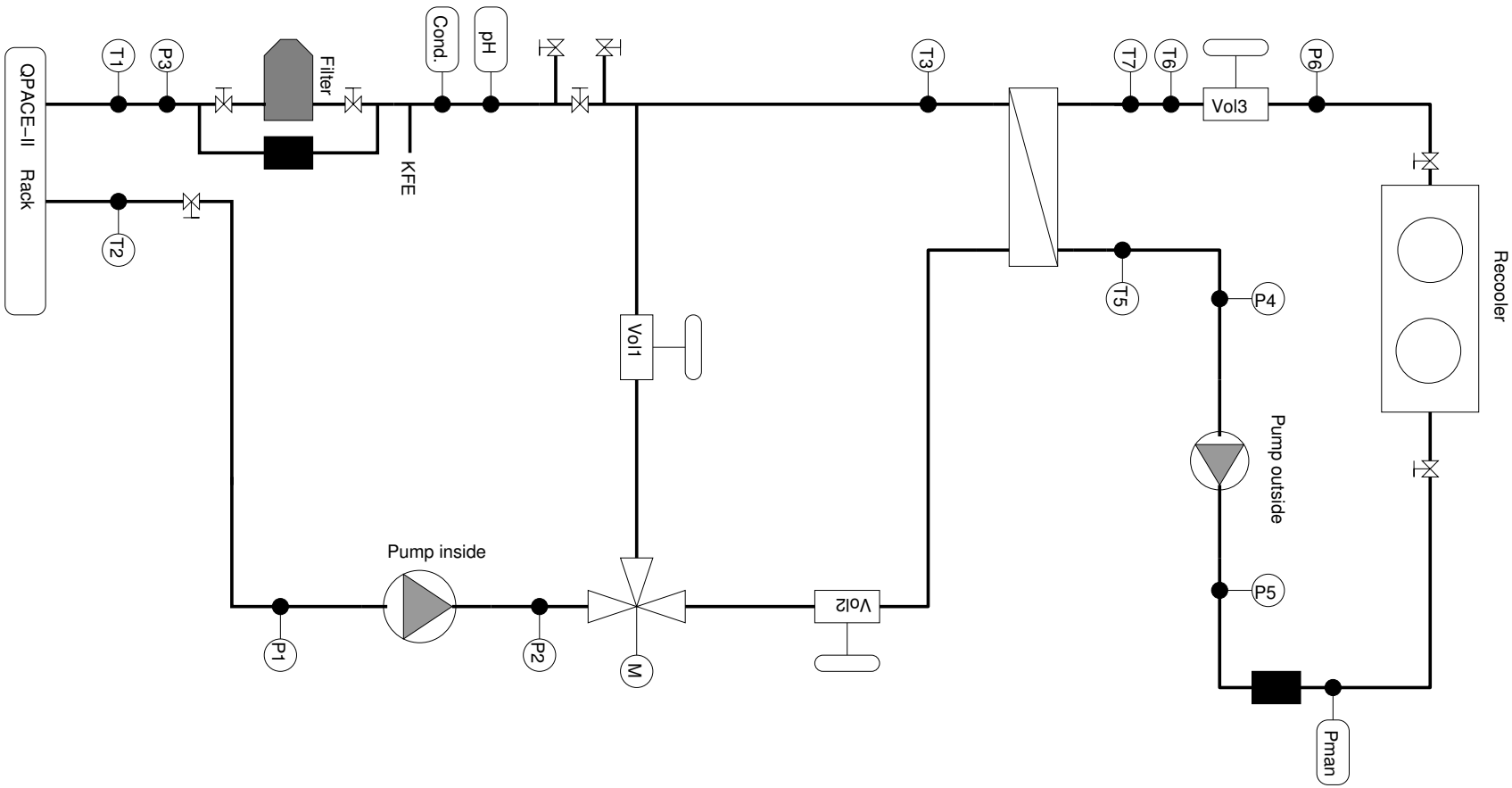
Low temperature difference between die and water is hard to achieve.

Go for warm water cooling.

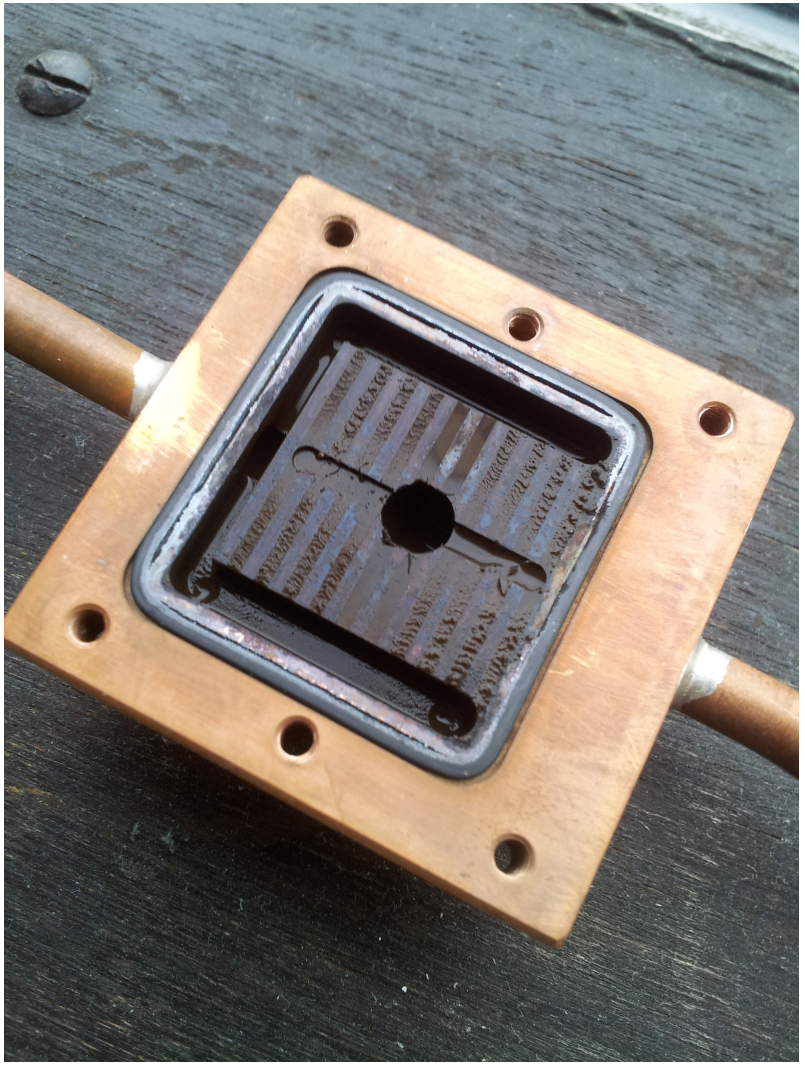
Lecture Notes in Computer Science Volume 7905, 2013, pp 383-394 iDataCool: HPC with Hot-Water Cooling and Energy Reuse

arXiv: 1309.4887

Cooling Infrastructure



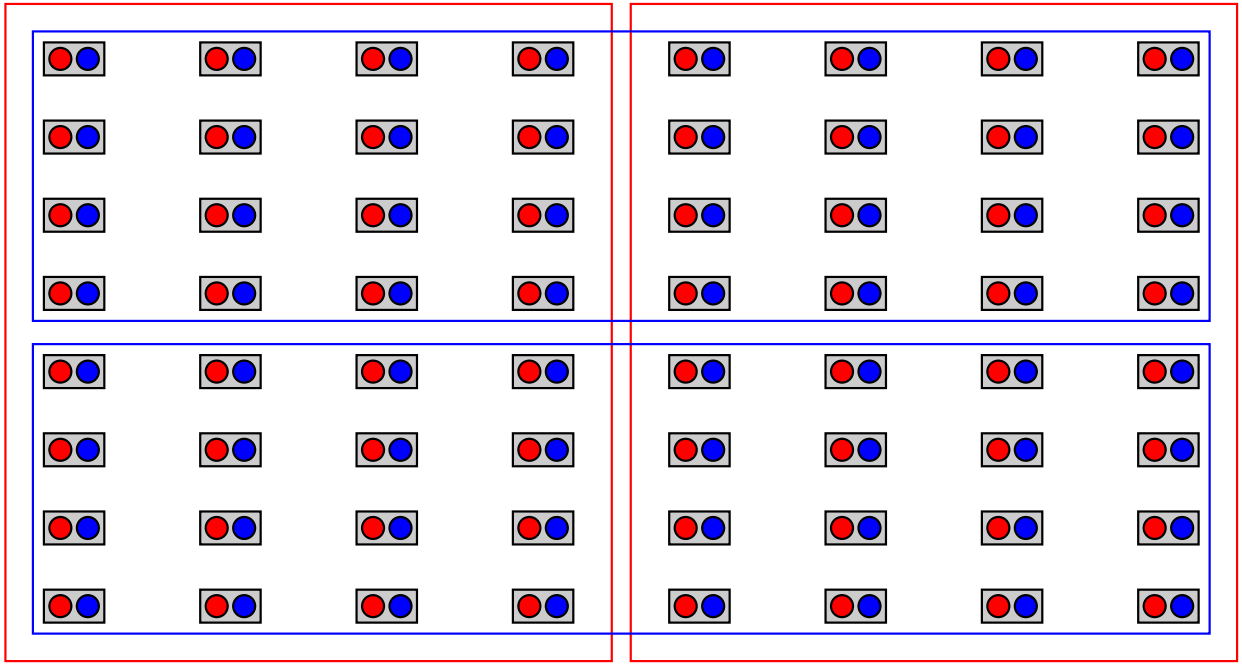
Lesson learned from idatacool: monitor everything!



Collect all sensor data in Cassandra

- pH
- conductivity
- pressure
- flow rate
- CPU temps
- Voltages
- ...

Network:
Infiniband with Hyper-Crossbar topology



Need 4 × 36 port switches. 4 ports per switch left, e.g., I/O (Lustre)

System is diskless. Netboot:

- PXE boot (syslinux)
- syslinux fetches kernel, inital ramdisk
- userland via NFS (NFS root filesystem)
- KNC specific software via NFS or Lustre

QPACE 2 and Domain Decomposition on the Intel Xeon Phi

Paul Arts, Jacques Bloch, Peter Georg, Benjamin Glaessle, Simon Heybrock, Yu Komatsubara, Robert Lohmayer, Simon Mages, Bernhard Mendl, Nils, Meyer Alessio Parcianello, Dirk Pleiter, Florian Rappl, Mauro Rossi, Stefan Solbrig, Giampietro Tecchiolli, Tilo Wettig, Gianpaolo Zanier

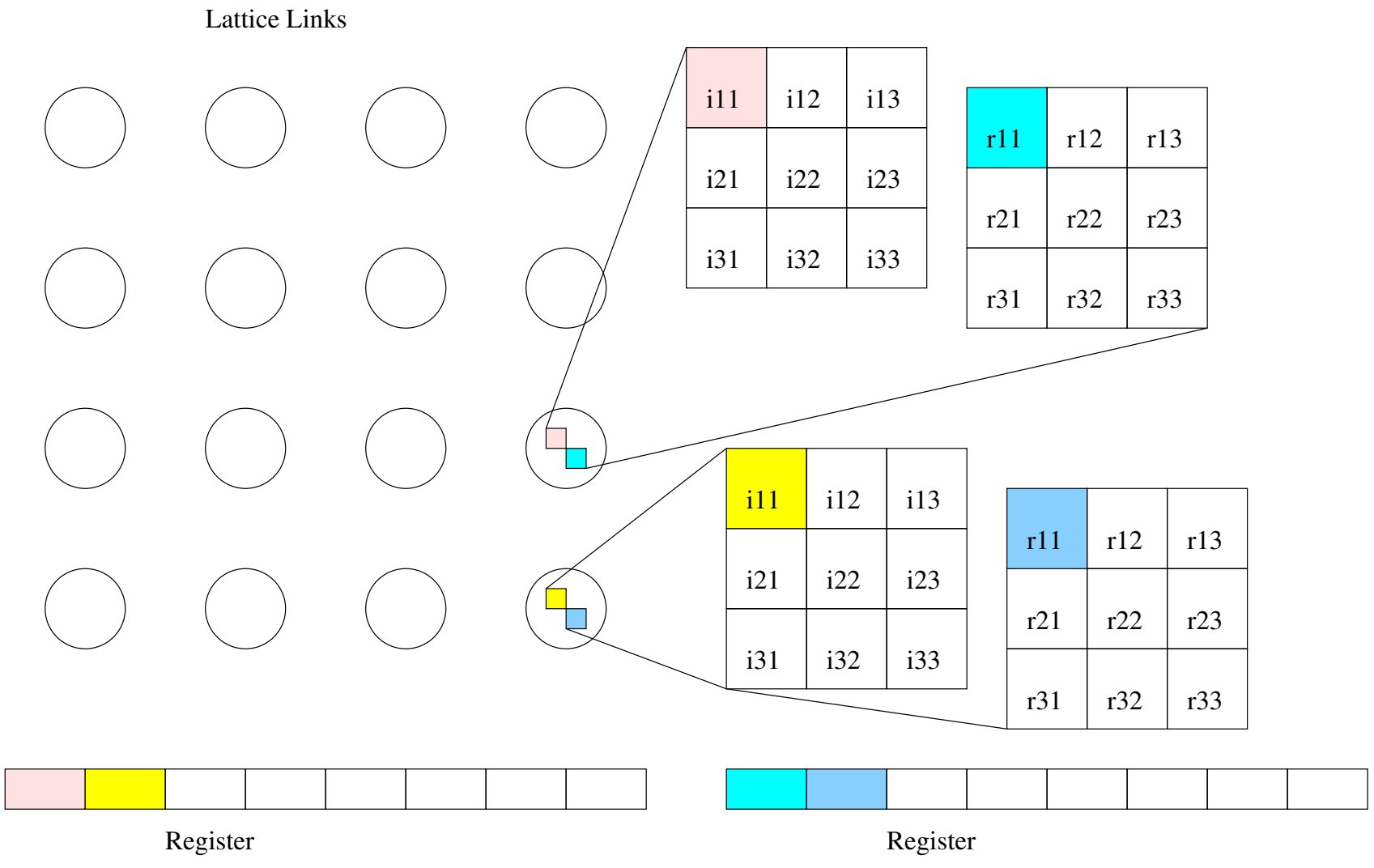
arXiv:1502.04025

Lattice QCD with Domain Decomposition on Intel Xeon Phi Co-Processors

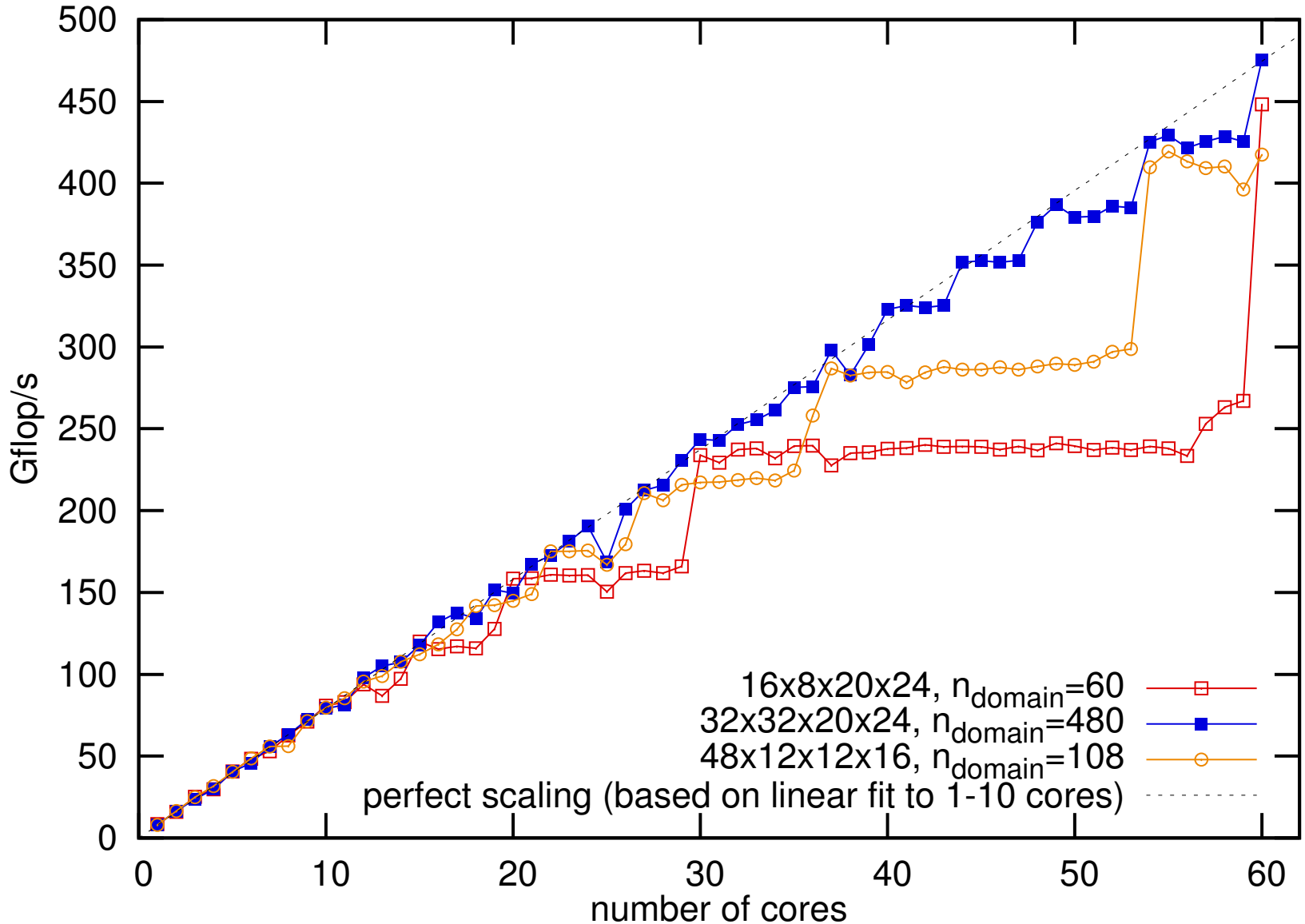
Simon Heybrock (Regensburg U.), Bálint Joó (Jefferson Lab), Dhiraj D. Kalamkar (Intel, Bangalore), Mikhail Smelyanskiy (Intel, Santa Clara), Karthikeyan Vaidyanathan (Intel, Bangalore), Tilo Wettig (Regensburg U.), Pradeep Dubey (Intel, Santa Clara)

arXiv:1412.2629

Coding — Site Fusing

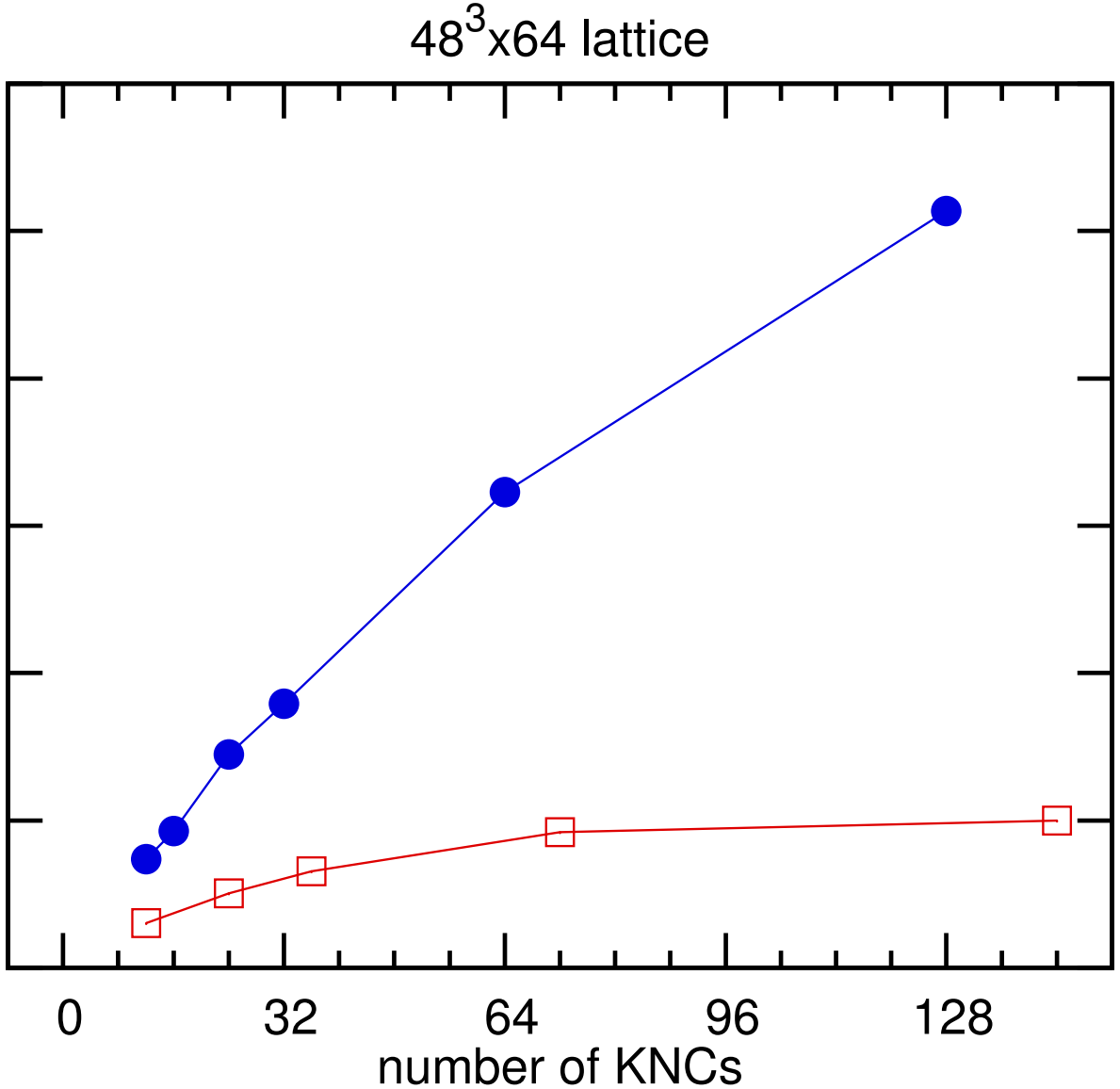


Coding — On Chip Scaling of DD preconditioner



from arXiv:1412.2629

Coding — Hard Scaling of DD preconditioner



from arXiv:1412.2629

Code

- Chroma runs
- Multigrid runs satisfactory (S.Heybrock).
- QDP threaded (J.Bloch)
- Rewrite of other parts would offer speedup compared to the one in QDP e.g., Wuppertal smearing.
- use SLURM as queueing system

Thank you for your attention!