DKRZ site news

The new system

Hendryk Bockelmann
Deutsches Klimarechenzentrum (DKRZ)

HLRE History

2002 HLRE-1
NECSX-6
Hurrikan
1,5 TeraFLOPS
60 TByte disk

100x compute
100x storage

2009 HLRE-2
IBM Power6
Blizzard
158 TeraFLOPS
6 PByte disk

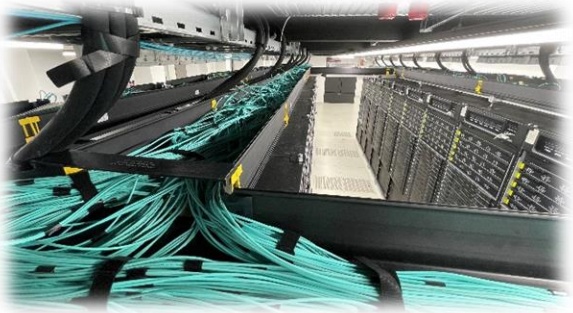
20x compute
8x storage

2015 HLRE-3
Atos bullx B700
Mistral
3 PetaFLOPS
50 PByte disk

2002 2004 2006 2008 2010 2012 2014 2016 2018 2020

2021 HLRE-4
Atos XH2000
Levante

Under Construction …



But since DKRZ is located in the middle of Hamburg:

- Energy supply up to 4MW at most
- Heat dissipation at the limit; use neighbouring roof space, manage noise emission
- Floor load at the limit

HPC by Atos XH2000 BullSequana

HPC system ‘Levante’	
CPU-nodes (AMD EPYC Milan 7763, 128 cores)	~2800 ca. 14,2 PFLOPS
GPU-nodes (as CPU nodes + 4*NVIDIA A100) Plus additional nodes in second phase	4 ca. 0,15 PFLOPS
Overall computing performance	ca. 16 PFLOPS ca. 5x improvement
Disk space (DDN)	120 PB ca. 2x improvement
Interconnect (NVIDIA Mellanox HDR)	200Gb/s ca. 4x improvement
Power consumption (incl. tape archive)	3,1 MW ca. 2x “improvement”

HSM by StrongLink (StrongBox Data Solutions)

HPC 5x higher production rate but just doubling of disk

$\Rightarrow$ New HSM system with key metrics

- Total capacity ready for 1.000 PB (1 EB)
- Annual throughput of 120 PB possible
- Max throughput rate 15 GB/s
- 1,2 PB disk cache
- 11 StrongLink nodes instead of 1 HPSS server for higher reliability
- Management of up to 2 billion objects + user defined MetaData

Room for Extension … What About GPUs?

Do we need a GPU-based system?

Does it pay off?

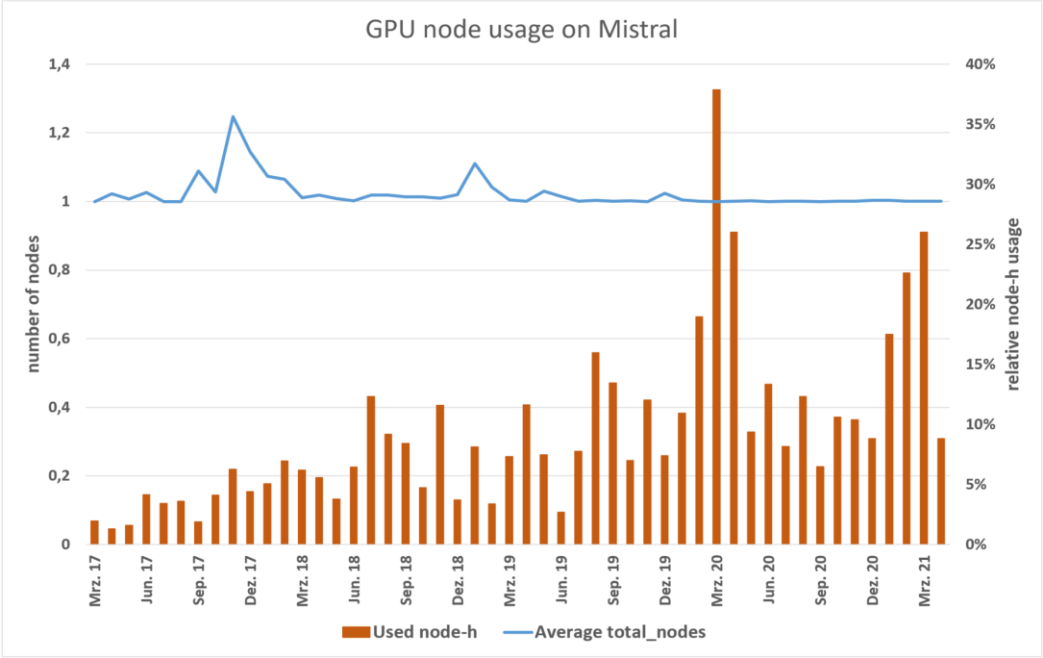
⇒ Cooperation between Atos, DKRZ and power users

Results:
The Good, the Bad and
the Ugly

What About GPUs ?

The Ugly (or definitely NO)

- Apart from ICON-A there is **no** other code of DKRZ users ready for GPUs yet …
- Already current system ‘Mistral’ (since 2015) has 20 GPU-nodes to prepare codes but …



What About GPUs ?

The Bad (or MAYBE)

- ML as a new use case ... still mostly just single GPU/node
- Programming paradigms ... hard to predict future
  - OpenACC, OpenMP, …
  - DSL or other frameworks like GridTools, kokkos, …
  - NVIDIA, AMD and Intel GPUs?

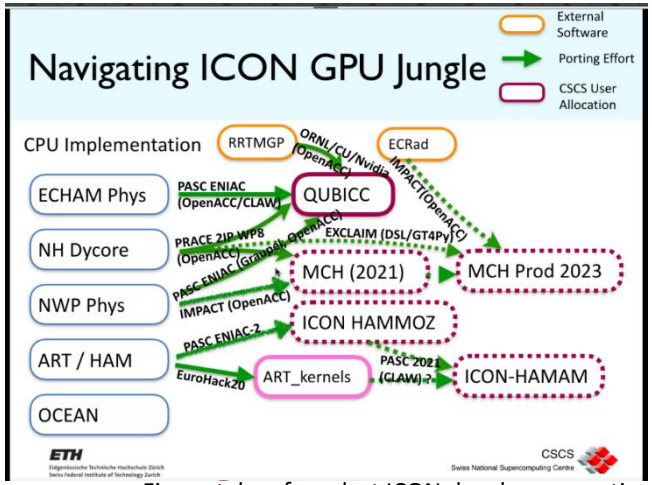
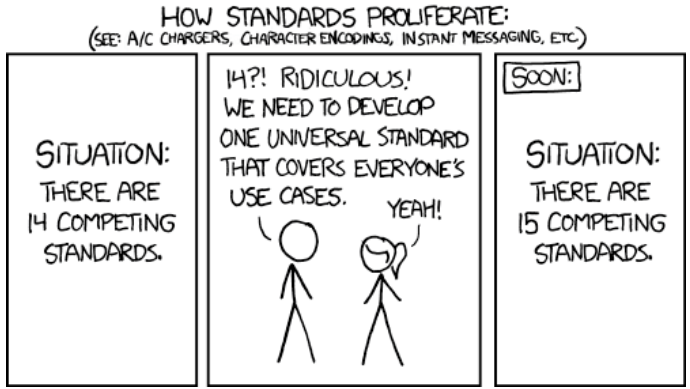


Figure taken from last ICON developer meeting

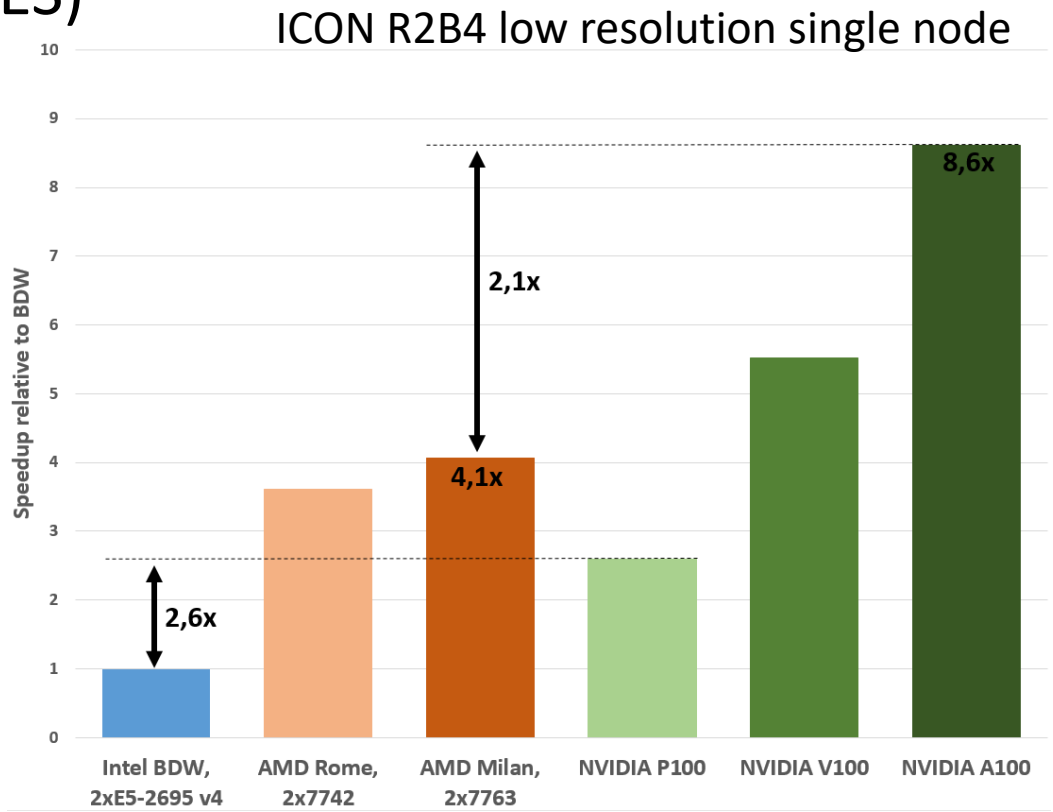


Taken from https://xkcd.com/927/

What About GPUs ?

The Good (or definitely YES)

- ICON atmosphere can be used
- Lower energy consumption of GPUs is promising



Technical Comparison CPU-GPU Nodes

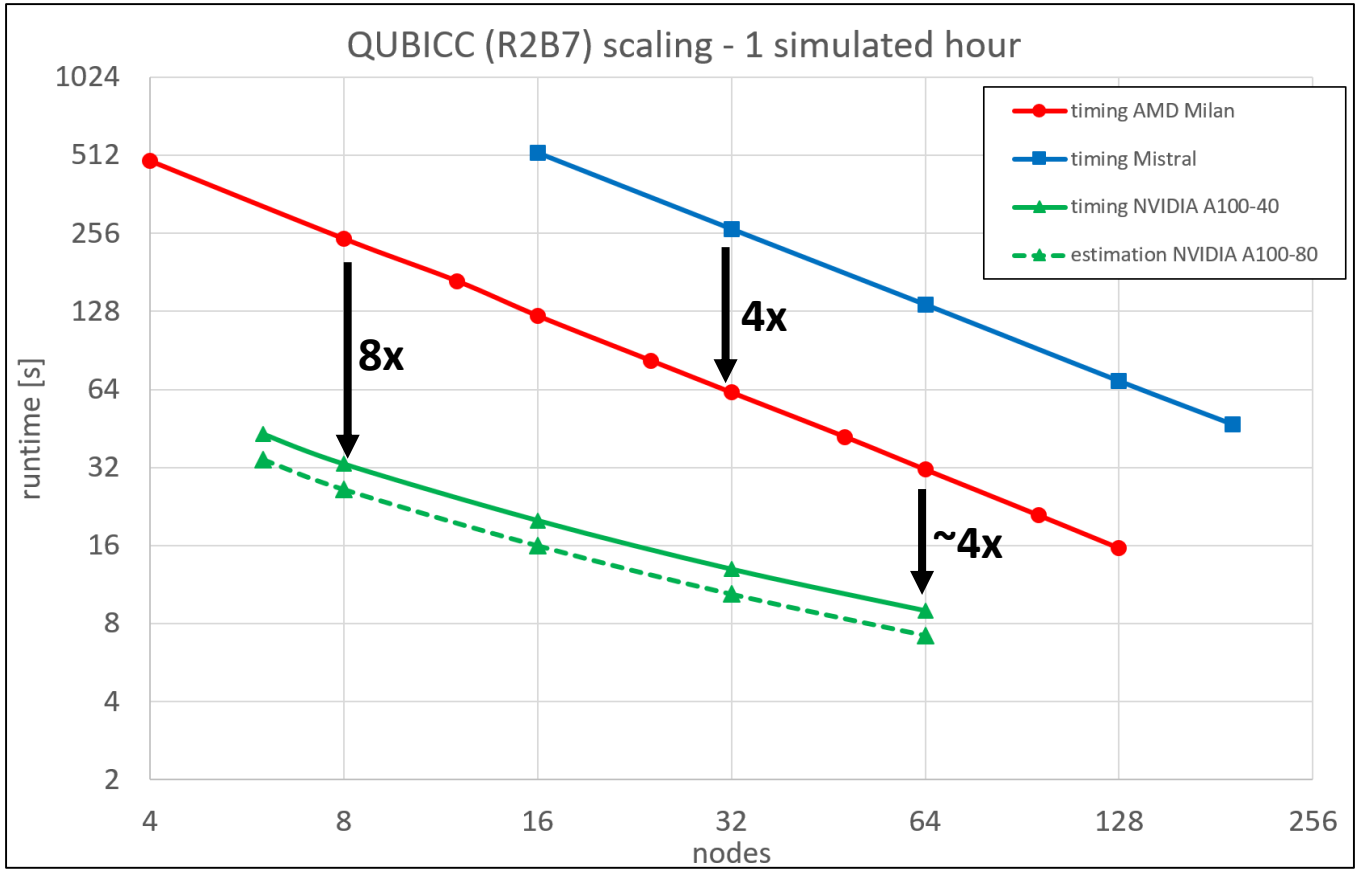
Based on well-known ICON-A experiments (only)

- Low resolution (R2B4, R2B7) SLAM and QUBICC for principle analysis and scalability
- High resolution (R2B9) QUBICC for throughput experiments

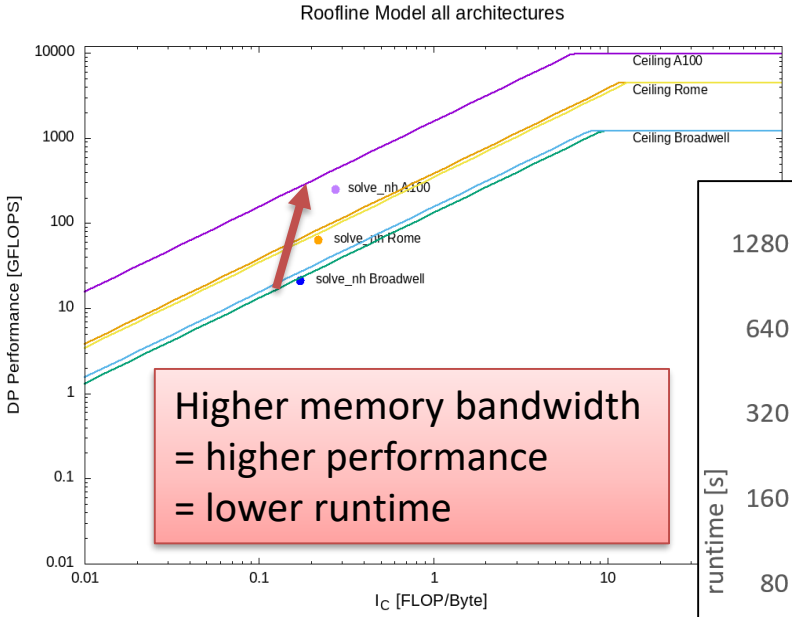
Systems compared: Mistral (Intel BDW), Piz Daint (NVIDIA P100), Spartan/Atos (AMD Milan), JUWELS booster (NVIDIA A100)

$\Rightarrow$ drawing a clear picture on what we can get out of phase2 as GPU or CPU

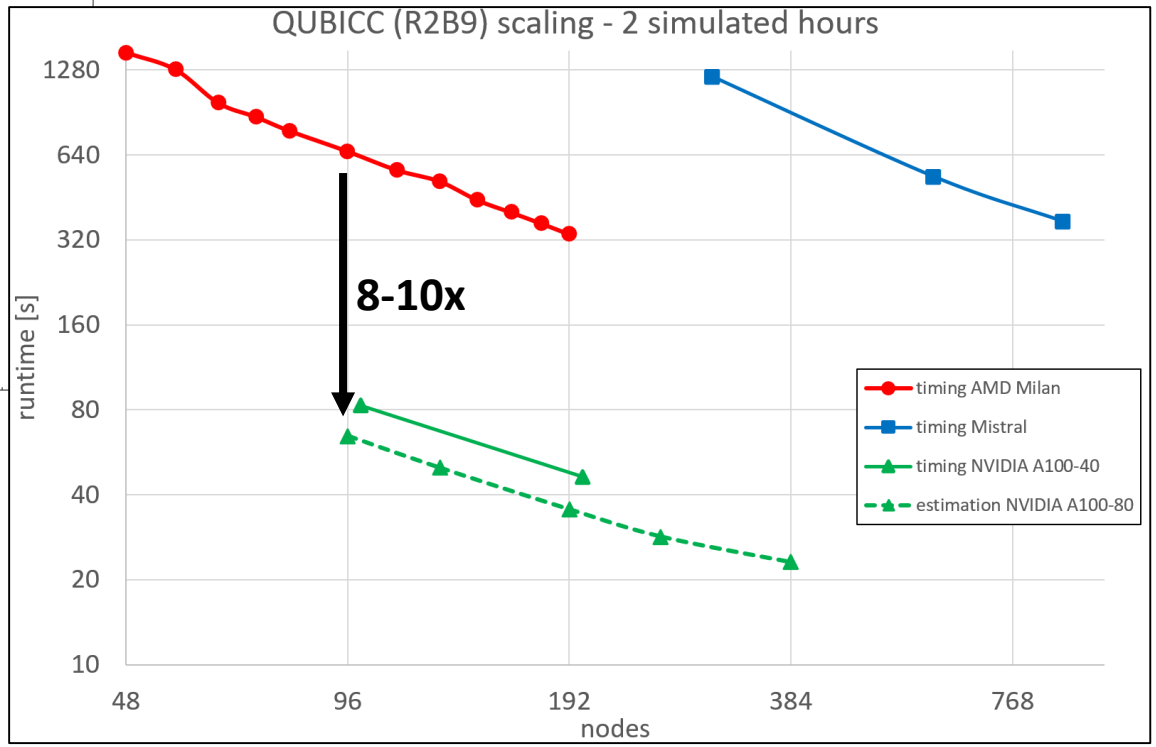
ICON Scaling Experiments



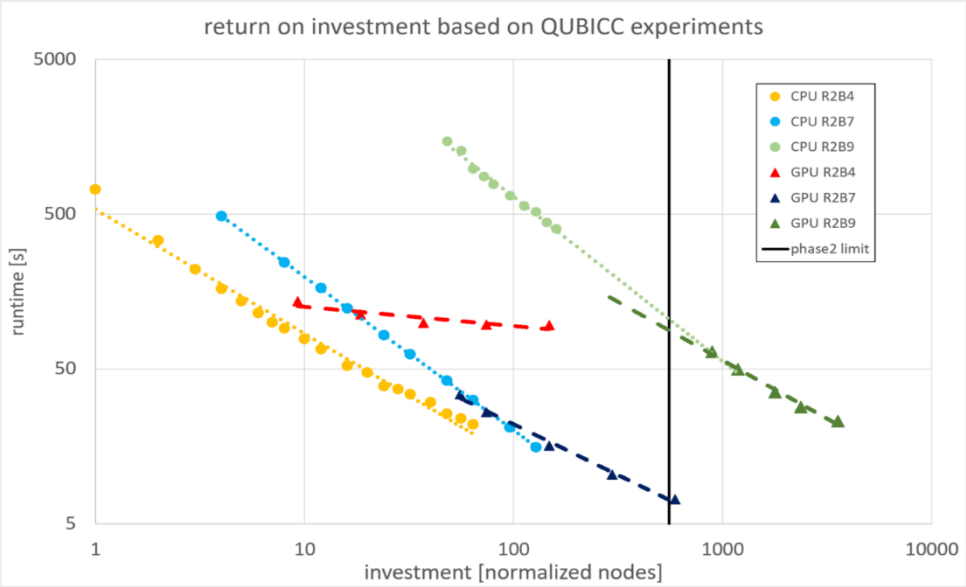
ICON Scaling Experiments



Higher memory bandwidth
= higher performance
= lower runtime



Return on Investment



General observation:

- GPUs use 40-50% less energy at same runtime if code saturates GPUs

Limitations:

- ICON R2B9 would **not** fit into 74 nodes with A100-40
- NVIDIA showed 39 nodes with A100-80 would work

	CPU	GPU A100-80	GPU A100-40
R2B7 SYPD	0.94 @192 nodes	0.95 @32 nodes	0.76 @32 nodes
#nodes for 1 SYPD	204	34	42
Energy for 1 SYPD	2942 kWh	1701 kWh	2125 kWh
SYPD on whole phase2	2.72	1.78	1.75

Outcome and Issues

- 9:1 exchange rate underlines the value of GPU nodes
  - Must be used from day 1, but code porting is high effort
- Majority of our users will benefit more from CPUs (capacity computing)
- NVIDIA A100-80 is well suited for high-res climate codes, if codes are ported (capability computing)
- Lower inlet temperature for GPU blades require active cooling
  - It remains to be seen whether full energy advantage is maintained

$\Rightarrow$ Compromise: Get one rack of GPU nodes for porting/preparing codes for Exascale EuroHPC system and ML