CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

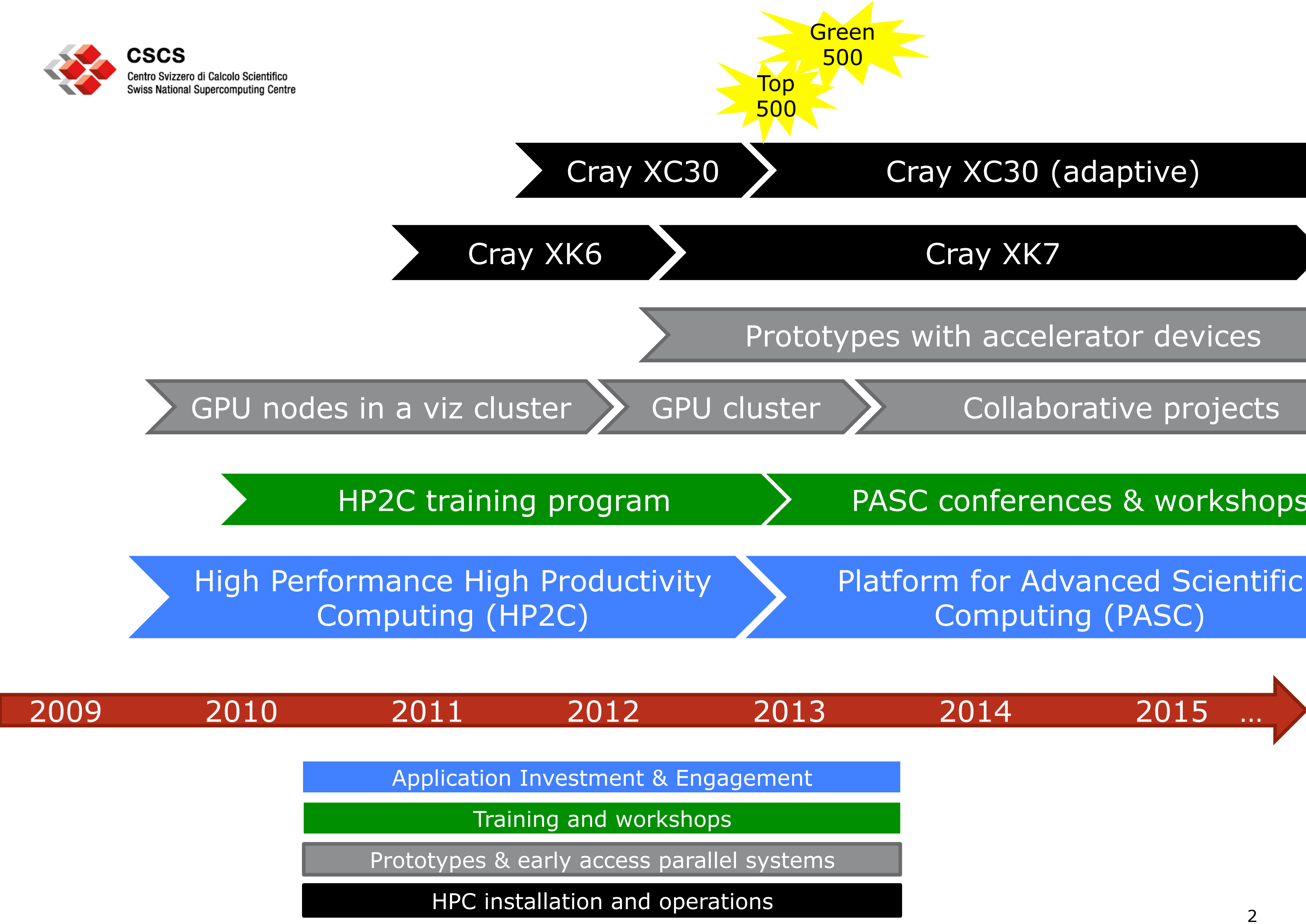
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

"Piz Daint:" Application driven co-design of a supercomputer based on Cray's adaptive system design

Sadaf Alam & Thomas Schulthess
CSCS & ETHzürich

CUG 2014

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre



Green 500

Top 500

Cray XC30

Cray XC30 (adaptive)

Cray XK6

Cray XK7

Prototypes with accelerator devices

GPU nodes in a viz cluster

GPU cluster

Collaborative projects

HP2C training program

PASC conferences & workshops

High Performance High Productivity Computing (HP2C)

Platform for Advanced Scientific Computing (PASC)

2009 2010 2011 2012 2013 2014 2015 …

Application Investment & Engagement

Training and workshops

Prototypes & early access parallel systems

HPC installation and operations

* Timelines & releases are not precise

GPU-enabled MPI & MPS

GPUDirect

GPUDirect-RDMA

OpenACC 1.0

OpenACC 2.0

OpenCL 1.0

OpenCL 1.1

OpenCL 1.2

OpenCL 2.0

Reduce code prototyping and deployment time on HPC systems

CUDA 2.x
CUDA 2.x
CUDA 2.x
CUDA 2.x

CUDA 3.x
CUDA 3.x
CUDA 3.x

CUDA 4.x
CUDA 4.x
CUDA 4.x

CUDA 5.x
CUDA 5.x

CUDA 6.x

Cray XK6

Cray XK7

Cray XC30 & hybrid XC30

X86 cluster with C2070, M2050, S1070

iDataPlex cluster M2090

Testbed with Kepler & Xeon Phi

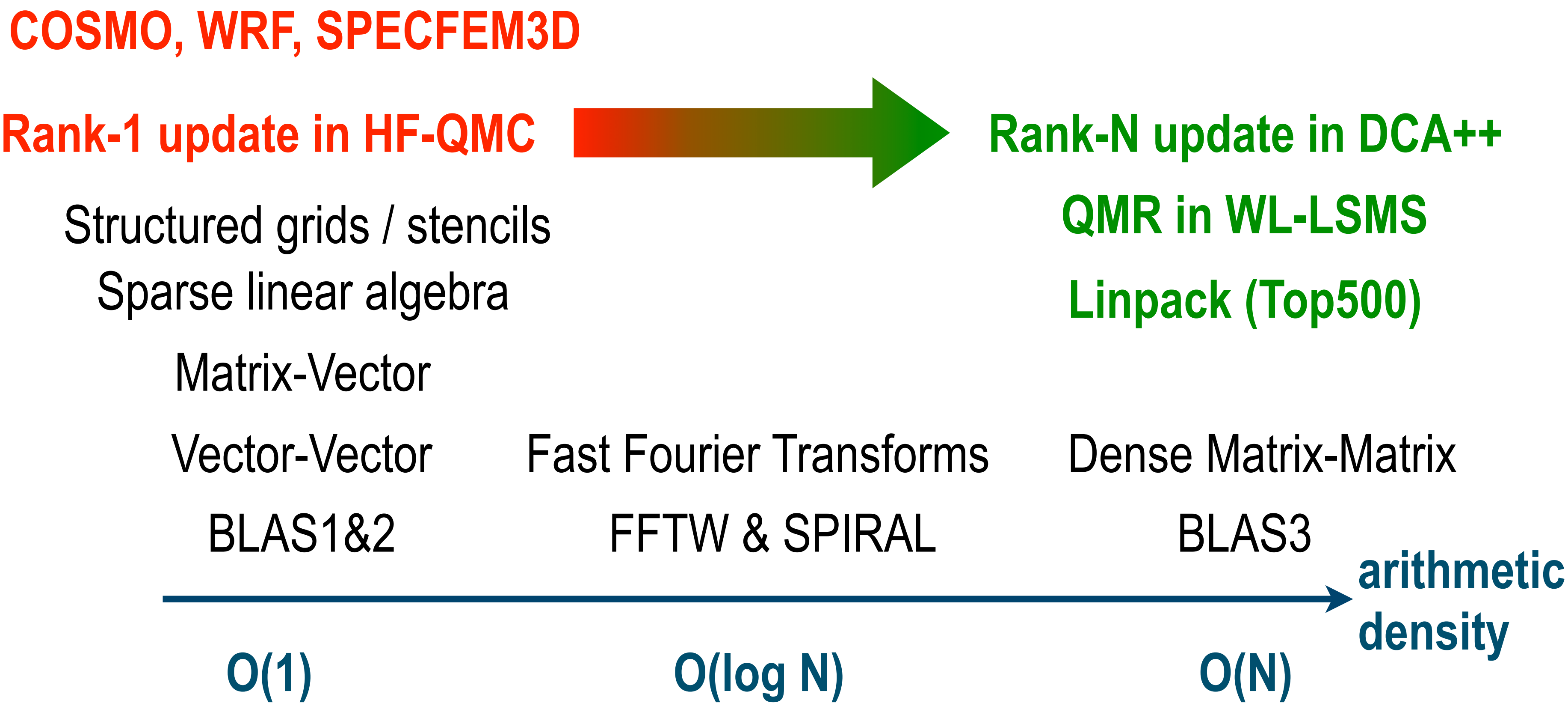
2009 2010 2011 2012 2013 2014 2015 …

Requirements analysis

Applications development and tuning

* Timelines & releases are not precise

Algorithmic motifs and their arithmetic intensity



COSMO, WRF, SPECFEM3D

Rank-1 update in HF-QMC

Rank-N update in DCA++

QMR in WL-LSMS

Linpack (Top500)

Structured grids / stencils

Sparse linear algebra

Matrix-Vector

Vector-Vector

BLAS1&2

Fast Fourier Transforms

FFTW & SPIRAL

Dense Matrix-Matrix

BLAS3

arithmetic density

O(1)

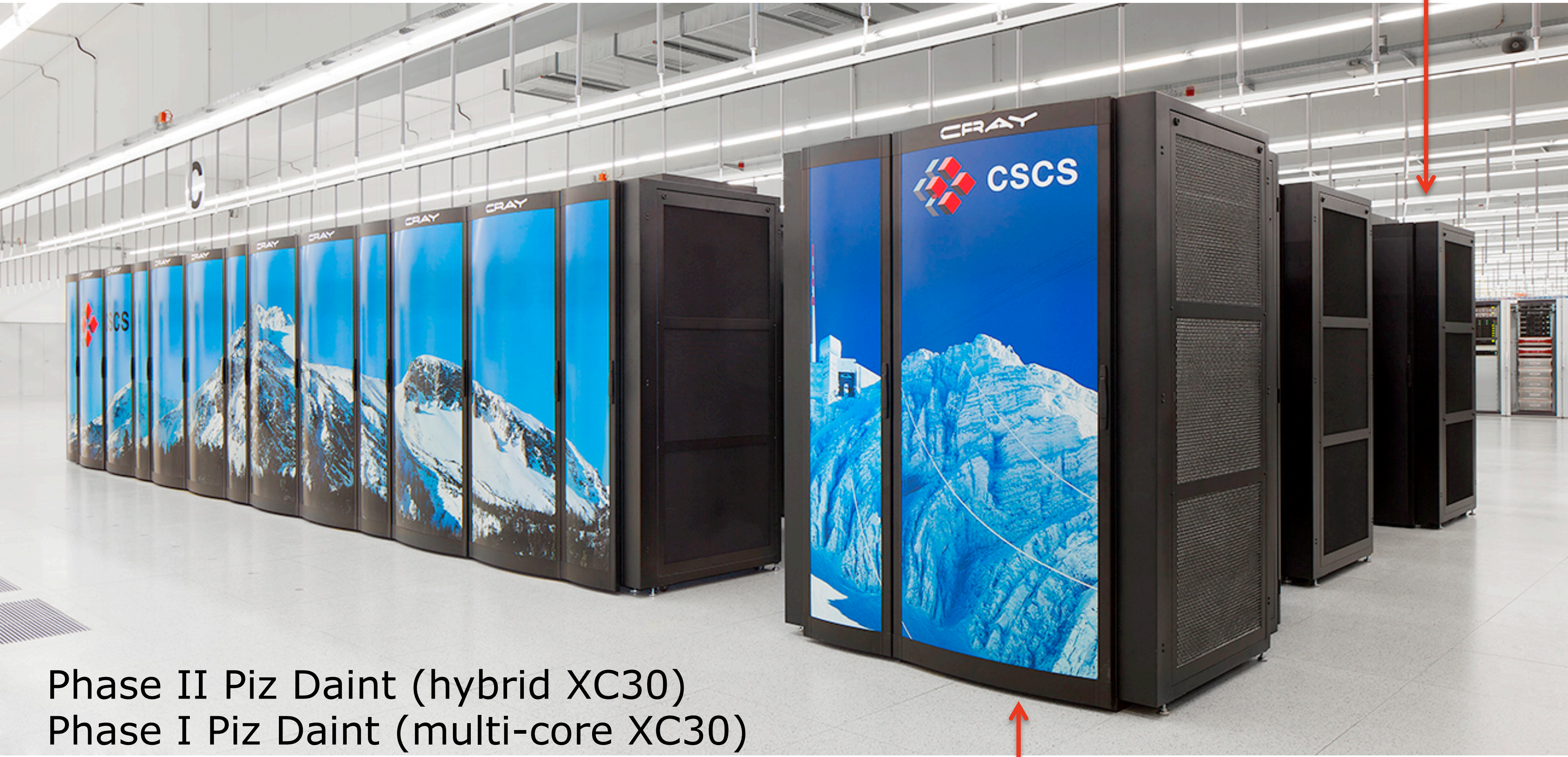
O(log N)

O(N)

Requirements Analysis

- **Compute and memory bandwidth**
  - Hybrid compute nodes
- **Network bandwidth**
  - Fully provisioned dragonfly on 28 cabinets
  - CP2K expose NW issues
- **GPU Enabled MPI & MPS**
- **Low jitter**

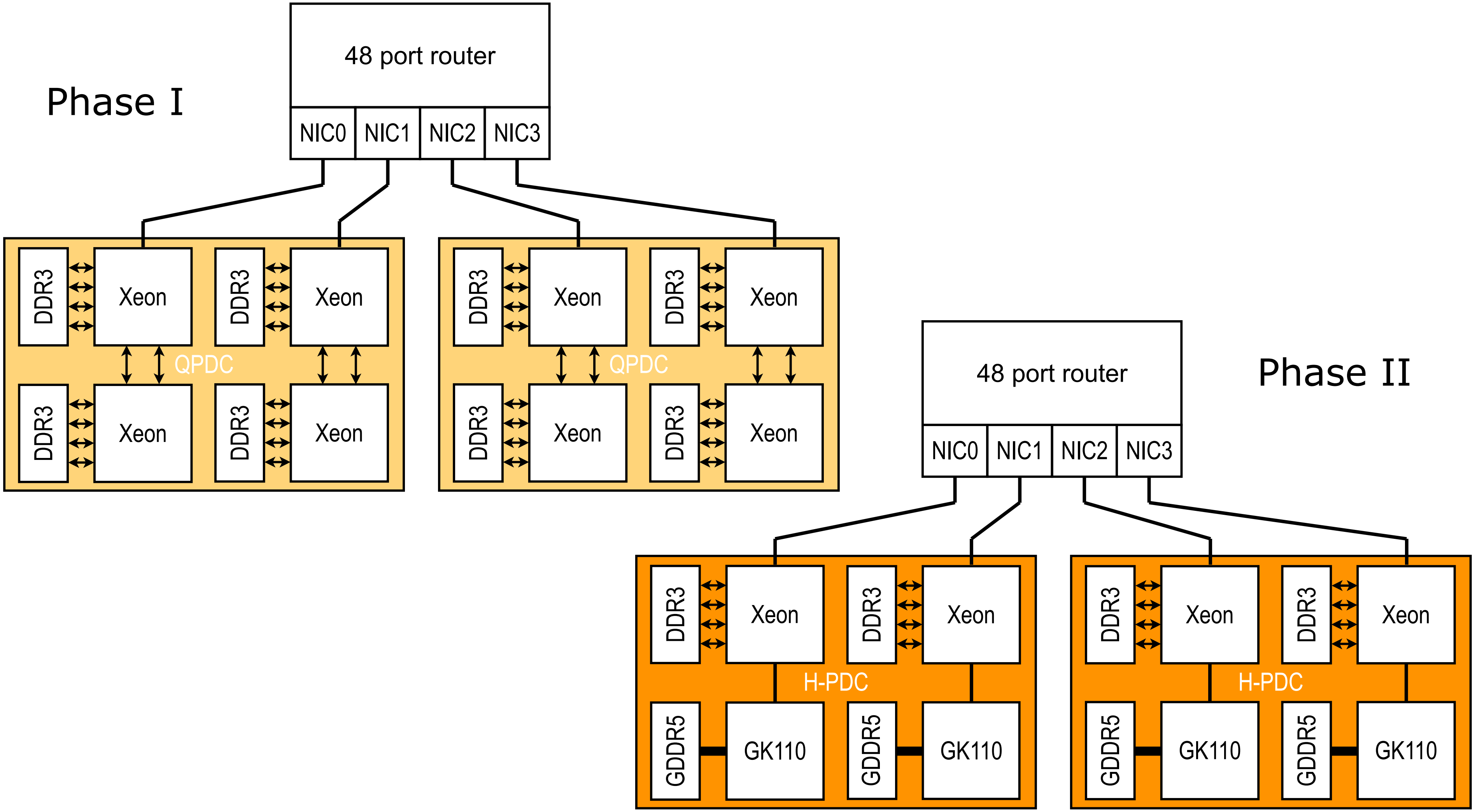
Third-row added
8 x 10 x10



Phase II Piz Daint (hybrid XC30)
Phase I Piz Daint (multi-core XC30)

TDS (santis)

Adaptive Cray XC30 Compute Node

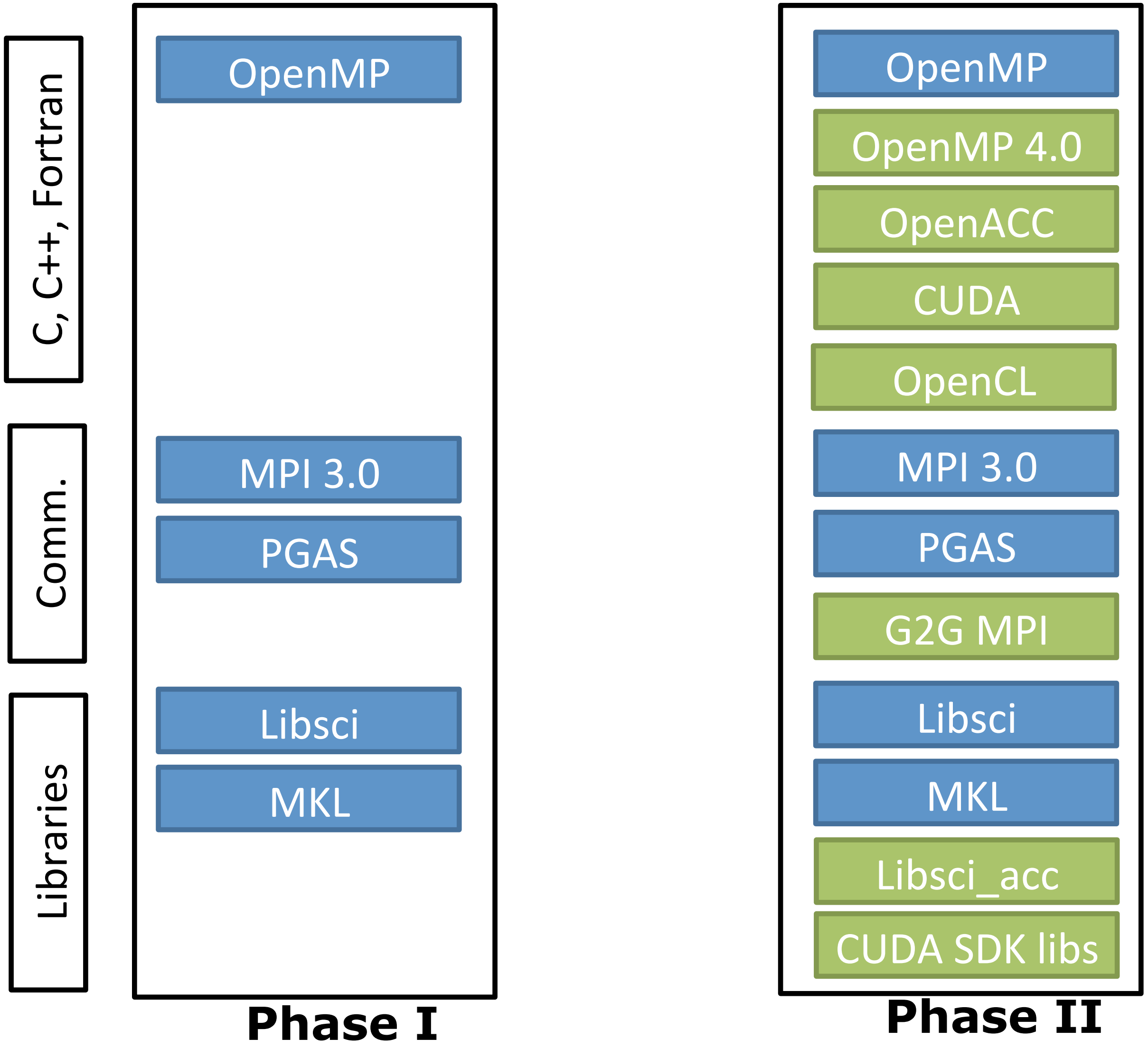


	Phase I	Phase II
Number of compute nodes	2,256	5,272
Peak system DP performance	0.75 Pflops	7.8 PFlops
CPU cores/sockets per node (Intel Xeon E5-2670)	16/2	8/1
DDR3-1600 memory per node	32 GBytes	32 GBytes
GPU SMX/devices per node (Nvidia K20x)	--	14/1
GDDR5 memory per node	--	6 GB (ECC off)
DP Gflops per node	332.8 Gflops (CPU)	166.4 Gflops (CPU) 1311 Gflops (GPU)
DDR3-1600 bandwidth per node	102.4 GB/s	51.2 GB/s
GDDR5 bandwidth per node	--	250 GB/s (ECC off)
Network & I/O interface	16x PCIe 3.0	16x PCIe 3.0
Network injection bandwidth per node	~ 10 GB/s	~ 10 GB/s

Node performance characteristics

	Phase I	Phase II
Number of cabinets	12	28
Number of groups	6	14
Number of Aries network & router chips	576	1344
Number of optical ports (max)	1440	3360
Number of optical ports (connected)	360	3276
Number of optical cables	180	1638
Bandwidth of optical cables	6750 GB/s	61425 GB/s
Bisection bandwidth	4050 GB/s	33075 GB/s
Point to point bandwidth	8.5-10 GB/s	8.5-10 GB/s
Global bandwidth per compute node	3 GB/s	11.6 GB/s

Network performance characteristics



C, C++, Fortran

- OpenMP

Comm.

- MPI 3.0
- PGAS

Libraries

- Libsci
- MKL

Phase I

C, C++, Fortran

- OpenMP
- OpenMP 4.0
- OpenACC
- CUDA
- OpenCL

Comm.

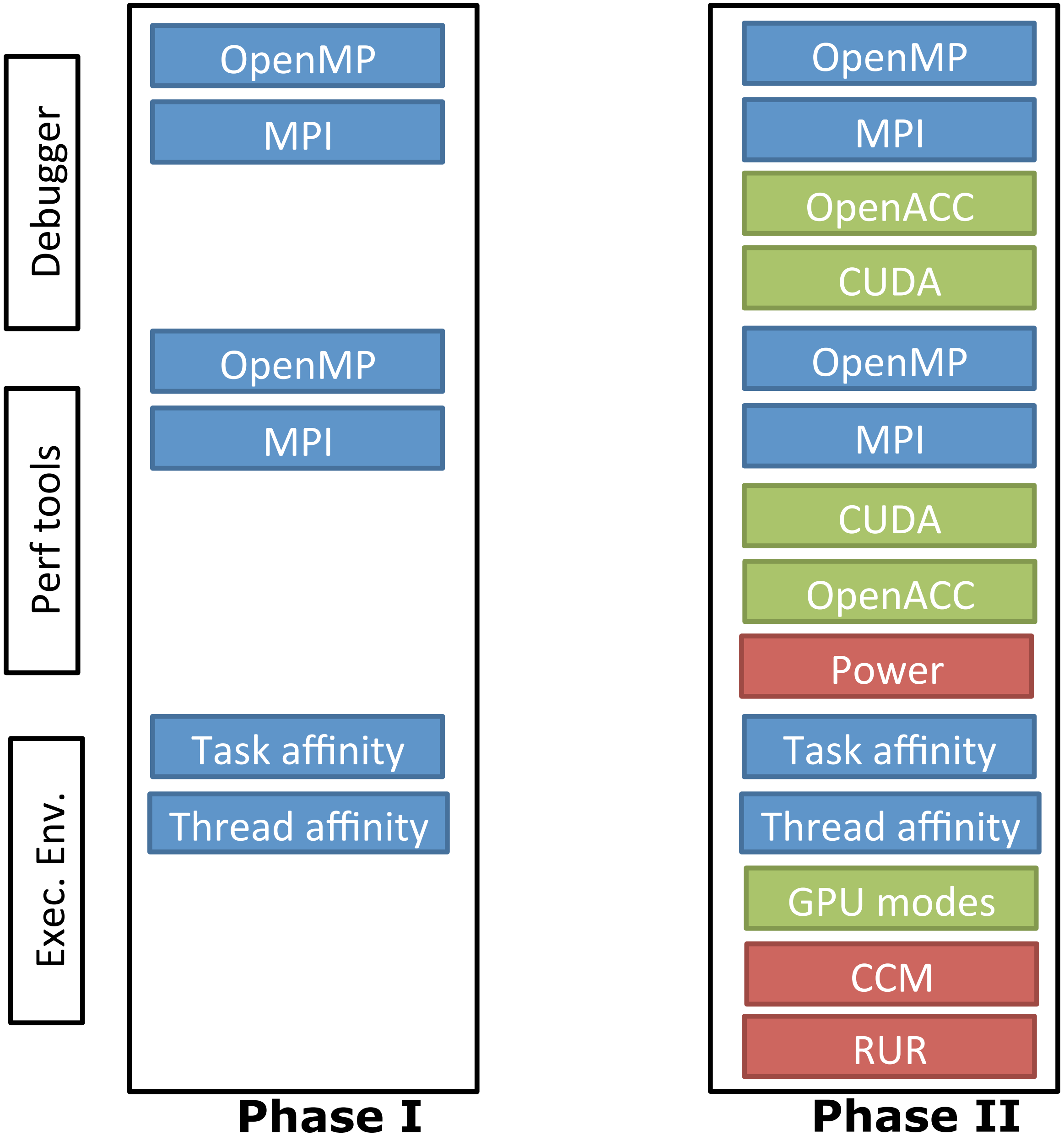
- MPI 3.0
- PGAS
- G2G MPI

Libraries

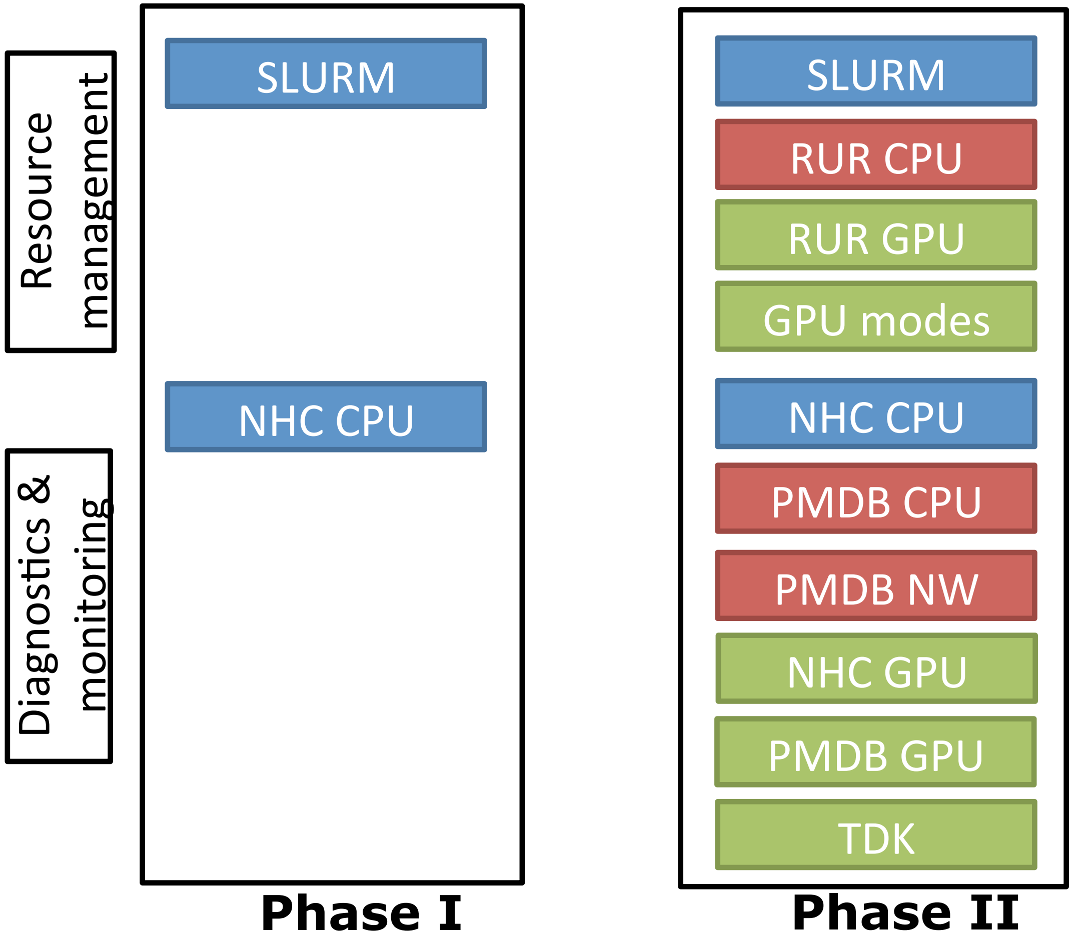
- Libsci
- MKL
- Libsci_acc
- CUDA SDK libs

Phase II

Co-designed Code Development Interfaces



Co-designed Tools and Execution Environment



Resource management

Diagnostics & monitoring

SLURM

NHC CPU

Phase I

SLURM

RUR CPU

RUR GPU

GPU modes

NHC CPU

PMDB CPU

PMDB NW

NHC GPU

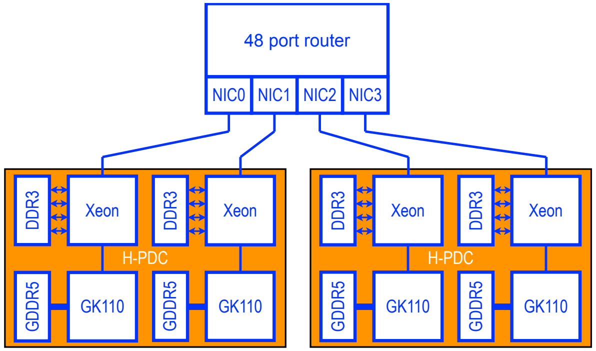
PMDB GPU

TDK

Phase II

Co-designed System Interfaces and Tools

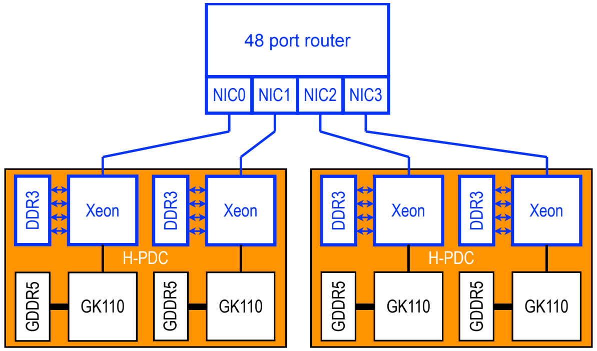
Adaptive Programming and Execution Models



Data Centric (I)

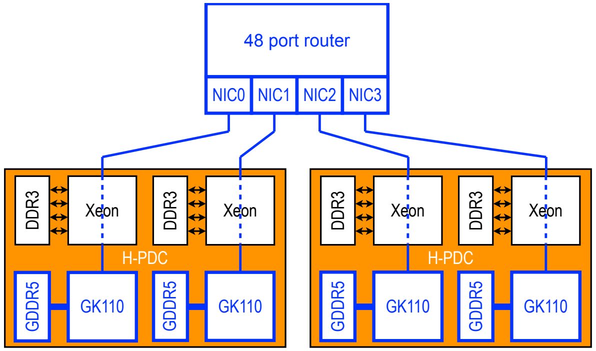
CPU DDR3
GPU GDDR5

Offload model



Data Centric (II)

CPU DDR3

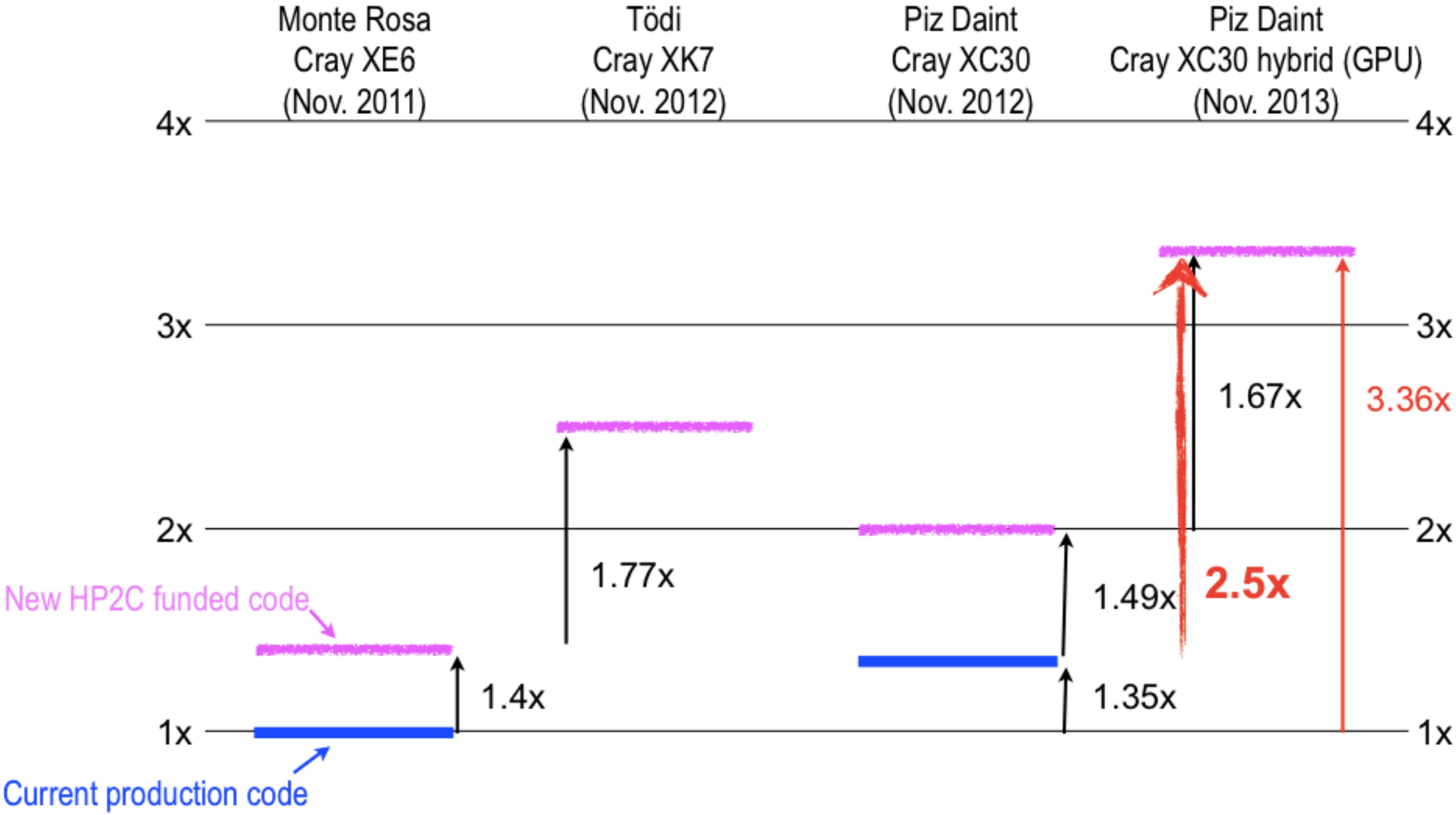


Data Centric (III)

GPU GDDR5
(+ CPU*)

GPUDirect-RDMA
GPU Enabled I/O

Speedup of COSMO-2 production problem – apples to apples comparison with 33h forecast of Meteo Swiss



Co-design potential for adaptive XC30

- **GPU for visualization**
  - CUG 2014 paper by Mark Klein (NCSA)
- **Extensions to programming environments**
  - Modular development tools (e.g. Clang-LLVM)
- **Improvements to the GPU diagnostics and monitoring**
  - Collaboration with Cray & Nvidia
- **Beyond GPUDirect … making GPU access other resources directly**

Thank you