ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CSCS
Swiss National Supercomputing Centre

Petascale Systems Challenges

Thomas C. Schulthess

SOS 13, Workshop on Distributed Computing
Hilton Head, South Carolina, March 9-12, 2009

What is a petascale system?

- A computer that can perform at least 10^{15} floating point operations per second - petaflop/s peak performance

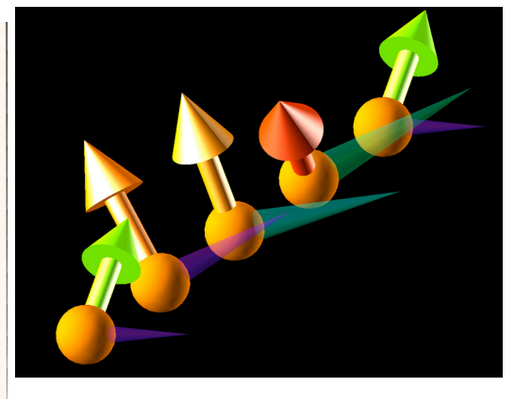
- A computer that can solve a large linear system of equations with an average rate of at least 10^{15} floating point operations - petaflop/s Linpack performance

- A computer that can solve (many) challenging scientific problems with an average rate of at least 10^{15} floating point operations - petaflop/s sustained application performance

From sustained gigaflop/s to teraflop/s to petaflop/s and beyond

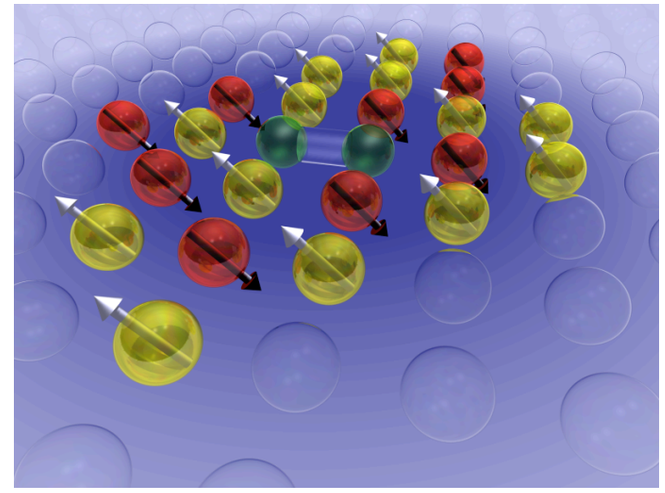
Evolution of the fastest sustained performance in real simulations

1.3 Gigaflop/s
Cray YMP
8 processors



1.02 Teraflop/s
Cray T_{3E}
$1.5\ 10^3$ processors

1.35 Petaflop/s
Cray XT5
$1.5\ 10^5$ processor cores

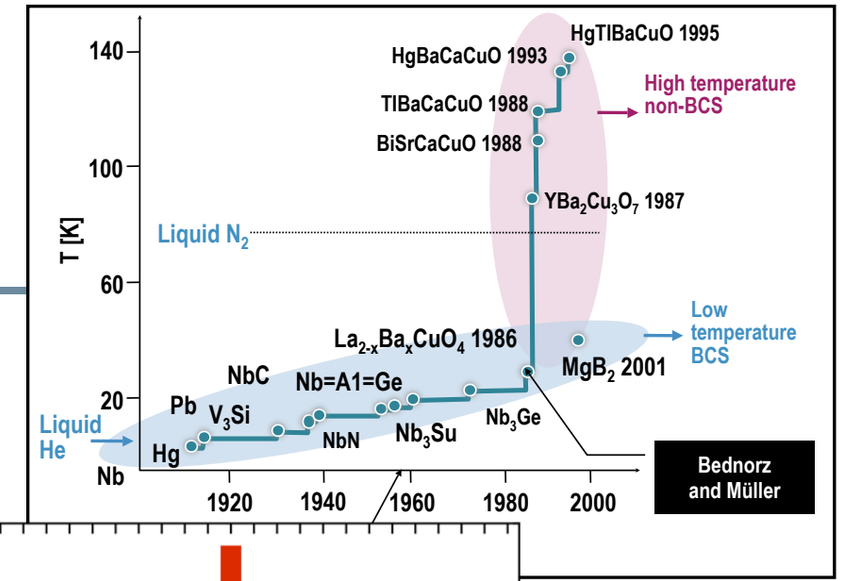


~1 Exaflop/s
~10^7 processing units

?

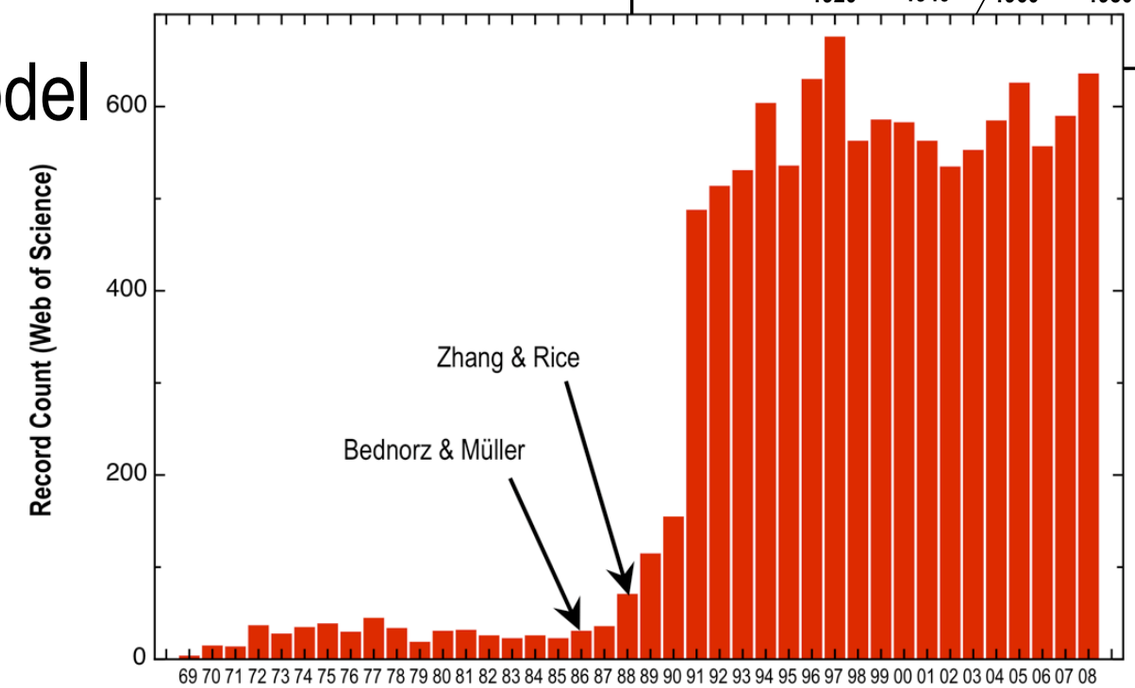
1989 1998 2008 2018

Pre-history



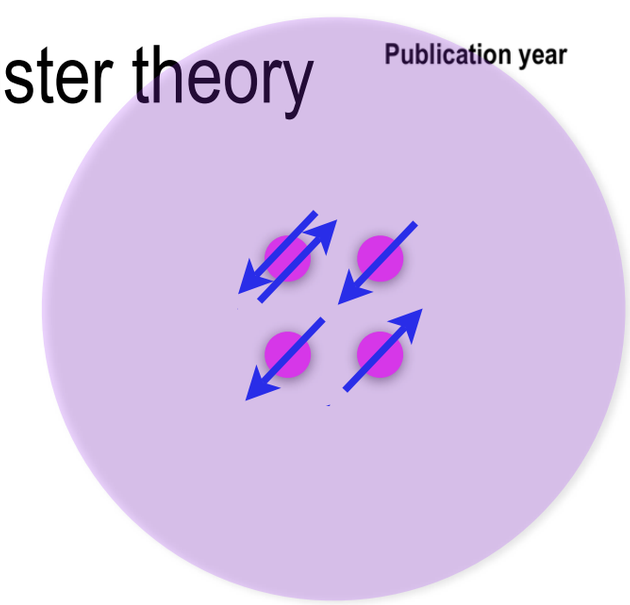
1986: Bednorz and Müller discover the cuprates

1988: Zhang & Rice motivate
single Band 2-D Hubbard model



>500 papers dealing with the problem appear each year
(counting only the ones recorded by Web of Science)

1998: Jarrell et al. formulate quantum cluster theory



ca. 2000: QMC/DCA method

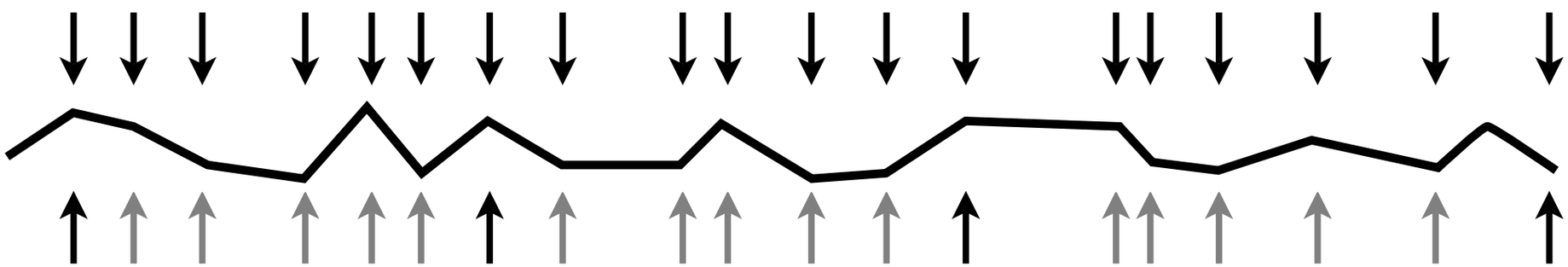
Systematic solution is now possible but computers are not powerful enough

Main milestones leading up to sustained petaflop/s with DCA++ code:

2004: Cray X1E powerful enough to solve model

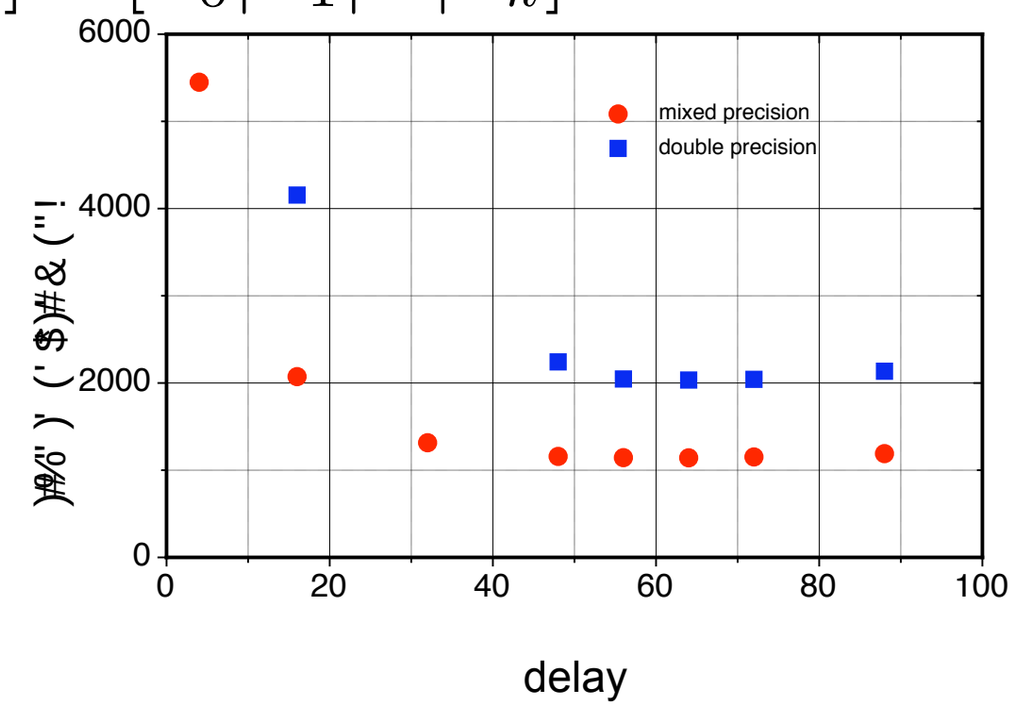
2005: First systematic solution of Hubbard model,
demonstrate it describes superconducting transition

$$\mathbf{G}_c(\{s_i,l\}_{k+1}) = \mathbf{G}_c(\{s_i,l\}_k) + \mathbf{a}_k \times \mathbf{b}_k^t$$

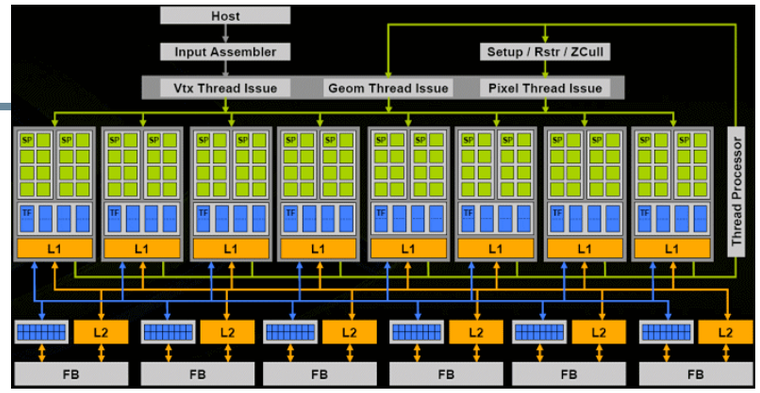


$$\mathbf{G}_c(\{s_i,l\}_{k+1}) = \mathbf{G}_c(\{s_i,l\}_0) + [\mathbf{a}_0|\mathbf{a}_1|...|\mathbf{a}_k] \times [\mathbf{b}_0|\mathbf{b}_1|...|\mathbf{b}_k]^t$$

2006/7: Delayed update algorithm improves
efficiency super-scalar processors
(sustaining ~50% of peak on Cray XT)



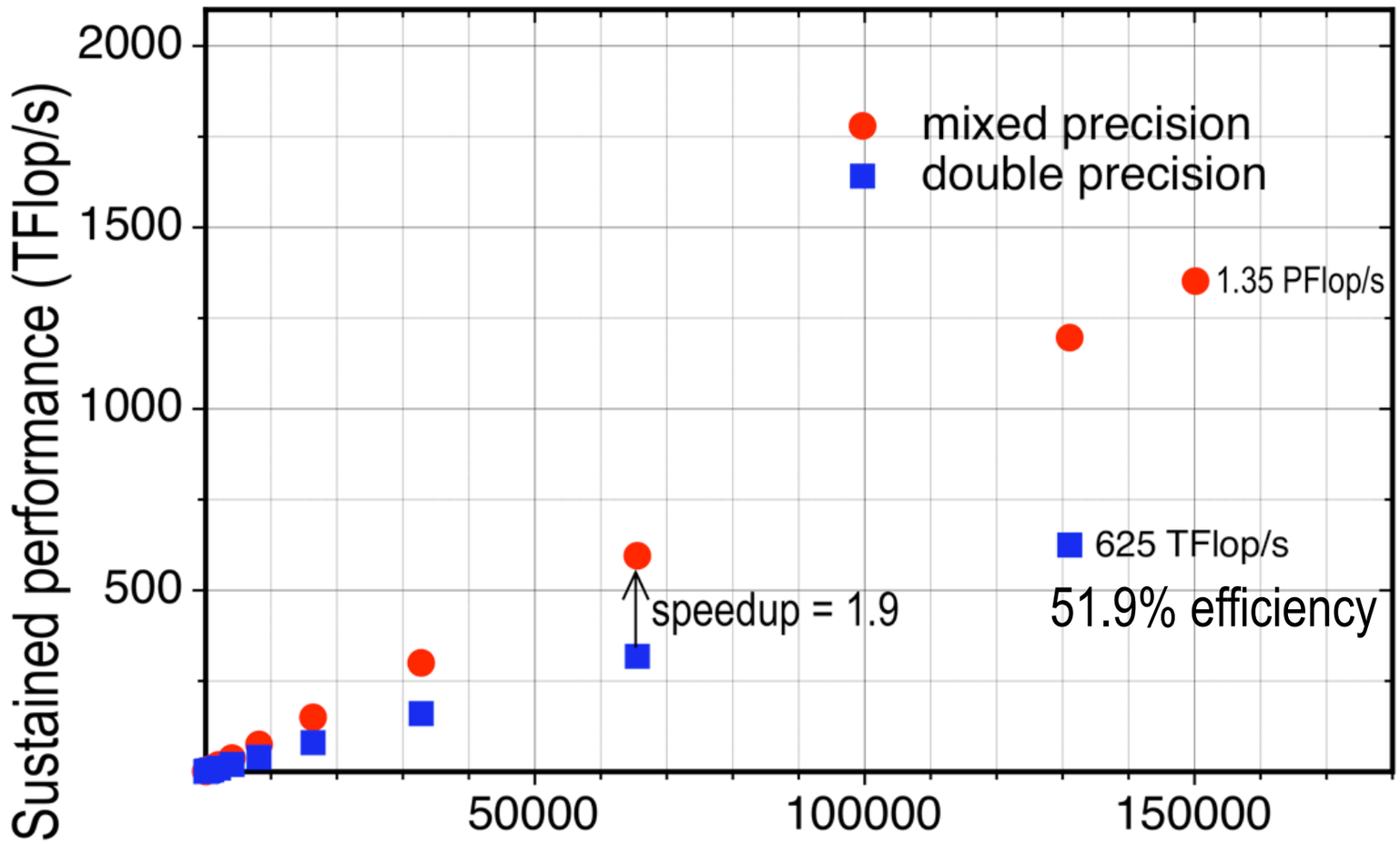
Main milestones leading up to sustained petaflop/s with DCA++ code (cont.):



01/2008: Experimenting with DCA++ on GPU motivates mixed precision algorithm

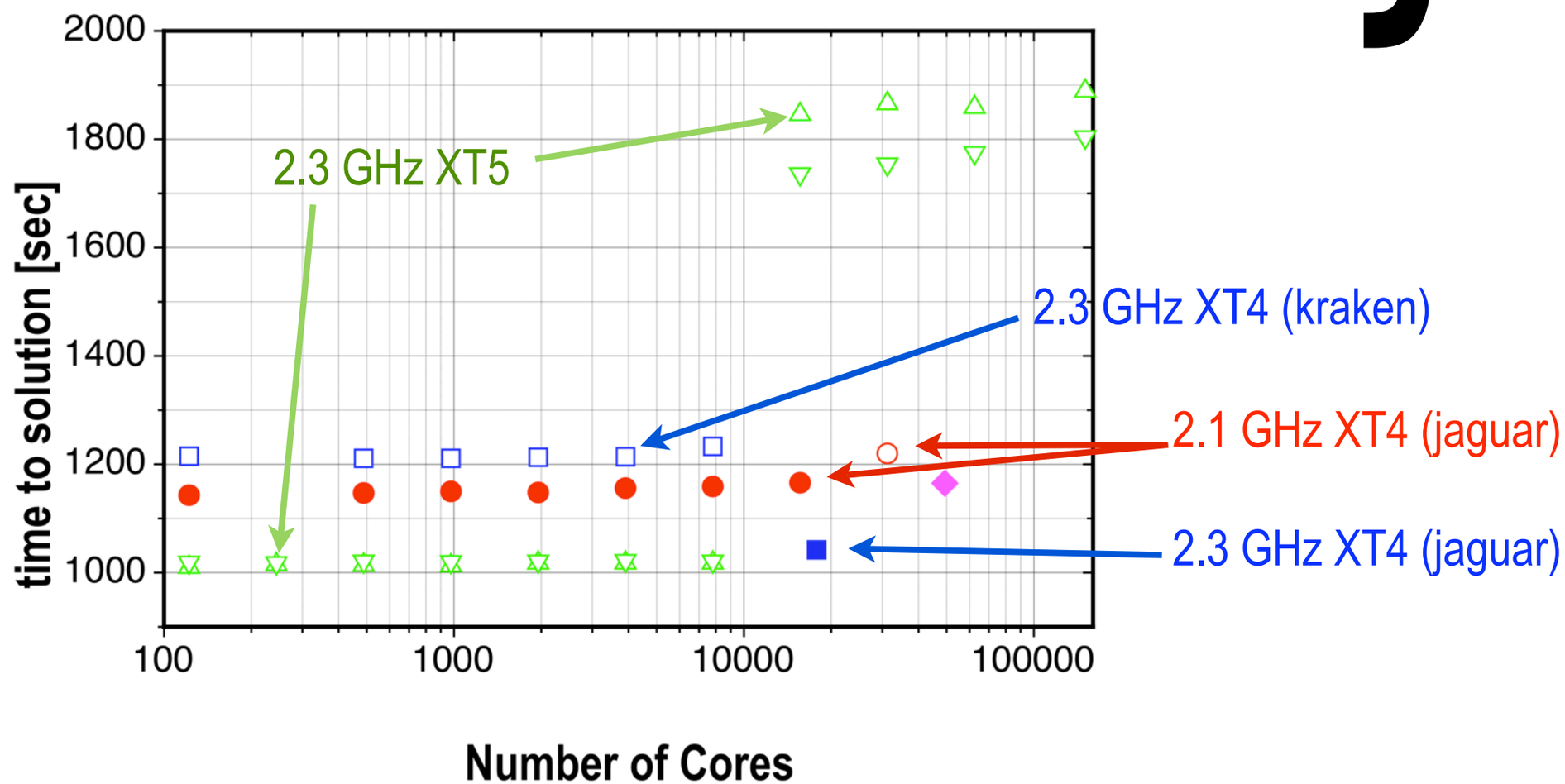
08/2008: After introduction of quad-core AMD, DCA++ sustained 409 TFlop/s on Cray XT4

11/2008: After introduction of Cray XT5, DCA++ sustained 1.35 PFlop/s



Three obvious hardware motivated challenges

- Performance on individual nodes
  - memory hierarchy, accelerators, multi-threaded
- Performance of the interconnect
  - latency, bandwidth, ... when 100,000+ cores communicate



- Resilience ... most runs take more than a few minutes!

But what should be even more obvious from our experience with petascale systems today

- The times when users developed applications and somebody a computer center helps porting the code to HPC systems are over!
- Solution methods, algorithms that optimally map to hardware, and application software are as much part of a petascale computing systems as are systems software, hardware, and operations.
- Collaborations between system designers, system operators, applied mathematicians, computer scientists, and domain scientists should motivate new computer systems and new approaches to solving challenging scientific problems.

What do we do about this?

New algorithm to enable 1+ PFlop/s sustained performance in simulations of disorder effects in high-T_c superconductors

G. Alvarez
M. S. Summers
D. E. Maxwell
M. Eisenbach
J. S. Meredith
J. M. Larkin
J. Levesque
T. A. Maier
P. R. C. Kent
E. F. D'Azevedo
T. C. Schulthess

D. Scalapino
M. Jarrell
J. Vetter
Trey White
staff at NCCS & Cray
many others

Computational resources:
NCCS @ ORNL

Funding:
ORNL-LDRD,
DOE-ASCR,
DOE-BES

OAK RIDGE National Laboratory

New algorithm to enable 1+ PFlop/s sustained performance in simulations of disorder effects in high-T_c superconductors

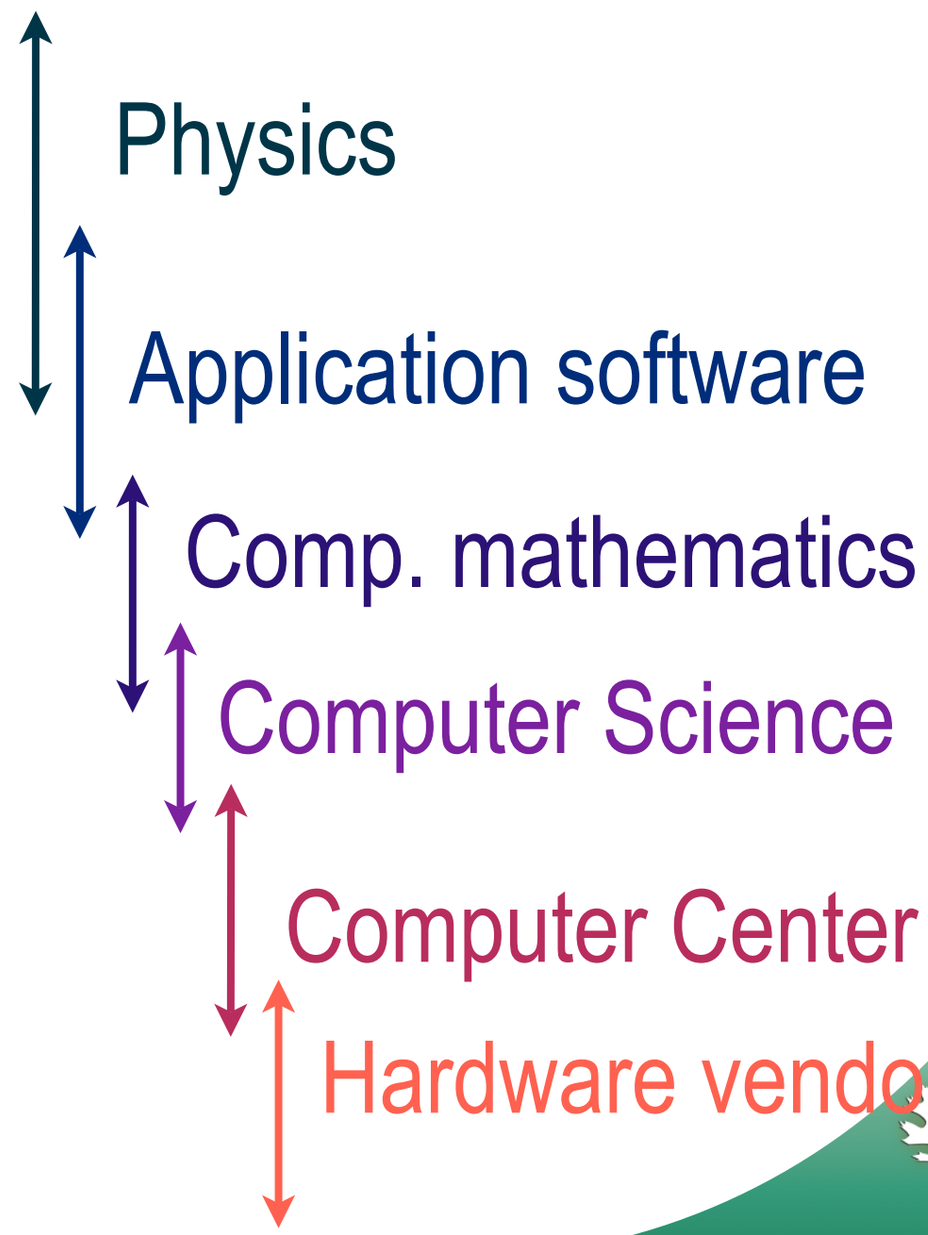
T. A. Maier
P. R. C. Kent
T. C. Schulthess
G. Alvarez
M. S. Summers
E. F. D'Azevedo
J. S. Meredith
M. Eisenbach
D. E. Maxwell
J. M. Larkin
J. Levesque

Models,
Methods,
& Implementation

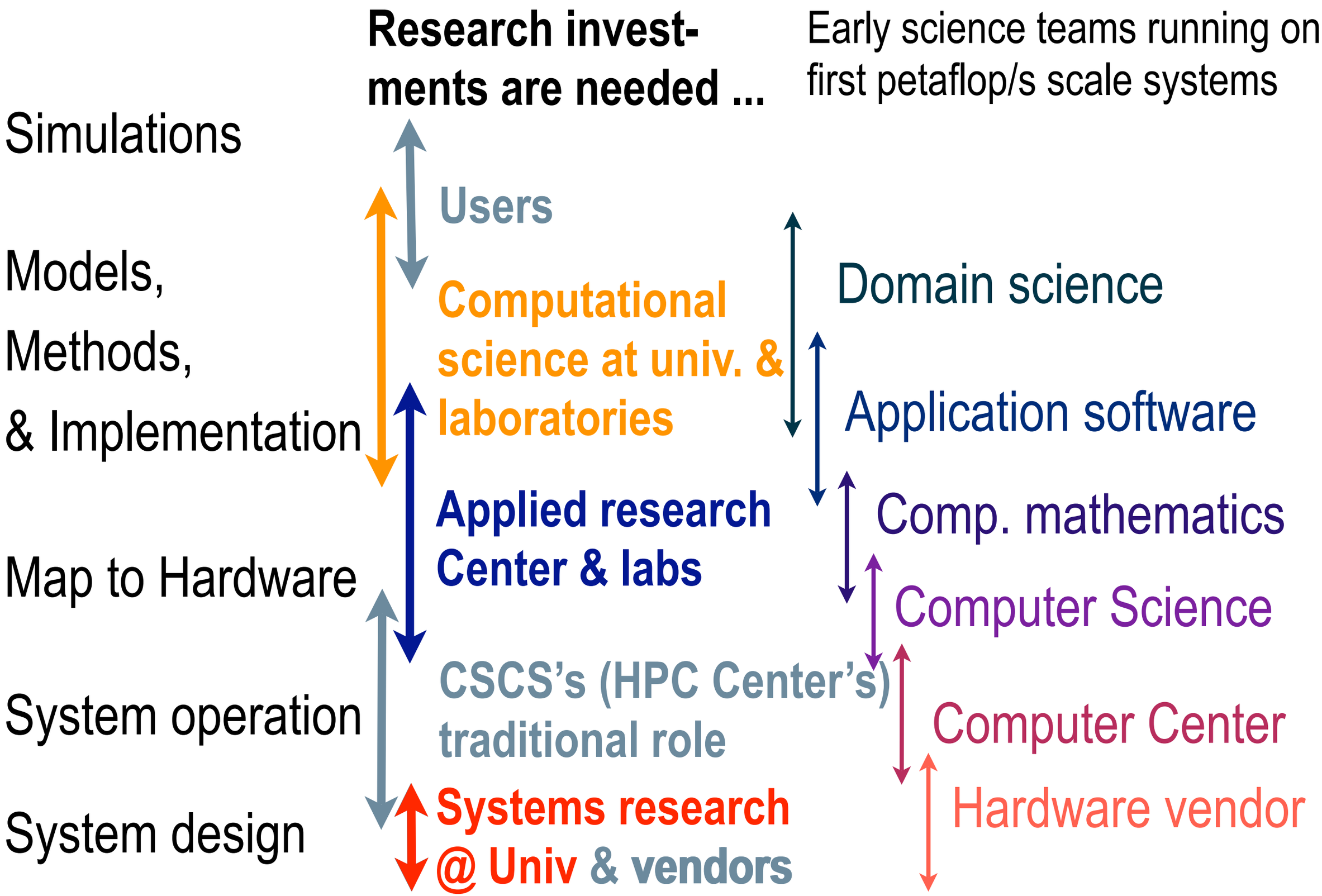
Map to Hardware

Operations

System design



Learning from the Oak Ridge experience: Covering all aspect of the simulation system



Elements of the High-Performance Computing and *Networking* (HPCN) initiative in Switzerland

- Develop HPC expertise and coordinate existing strengths in computational sciences (2009-2012)
  - Core program for computational math & applied computer science at CSCS and University of Lugano (~15 staff / faculty)
  - 10-15 projects high-risk / high-payoff projects (like DCA++) lead by existing computational science projects & embedded postdoctoral fellows / graduate students (liaison to core program)
  - Include HPC in CSE curricula at universities
  - Work with vendors / industry / US natl. labs to introduce prototype hardware
- Upgrade building infrastructure
  - CSCS in new building near University of Lugano by 2012
- Investments in hardware
  - Phase I: upgrade Cray XT system to maximum possible with current building infrastructure - goal is to get to several 10^4 cores quickly
  - Phase II: operate new petascale systems in 2012 - new procurement based on established needs of user network

Questions / Comments?