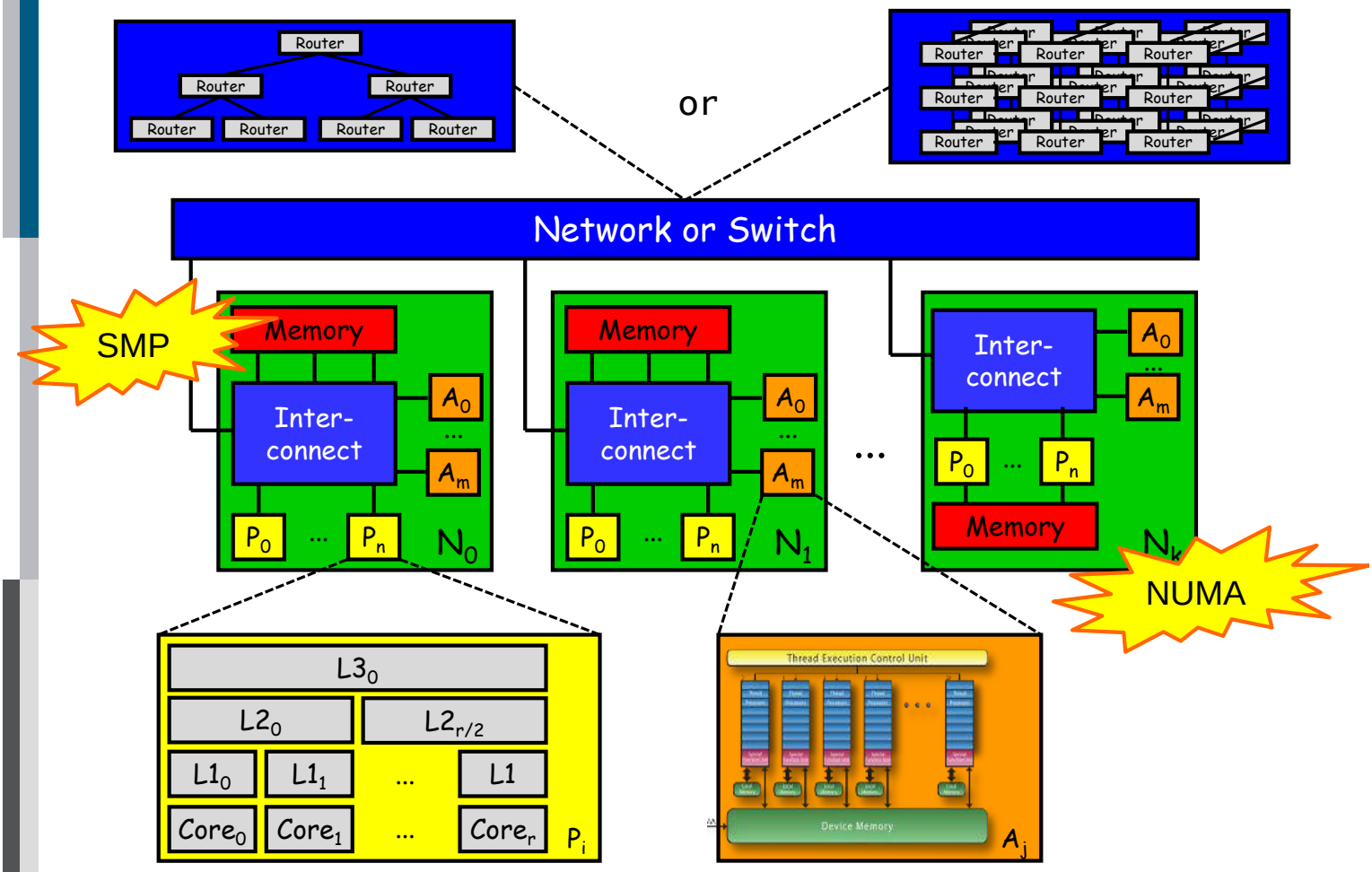
Scalasca: A Scalable Portable Integrated Performance Measurement and Analysis Toolset

Bernd Mohr

Parallel Architectures: State of the Art



Exascale Performance Challenges

JÜLICH FORSCHUNGSZENTRUM

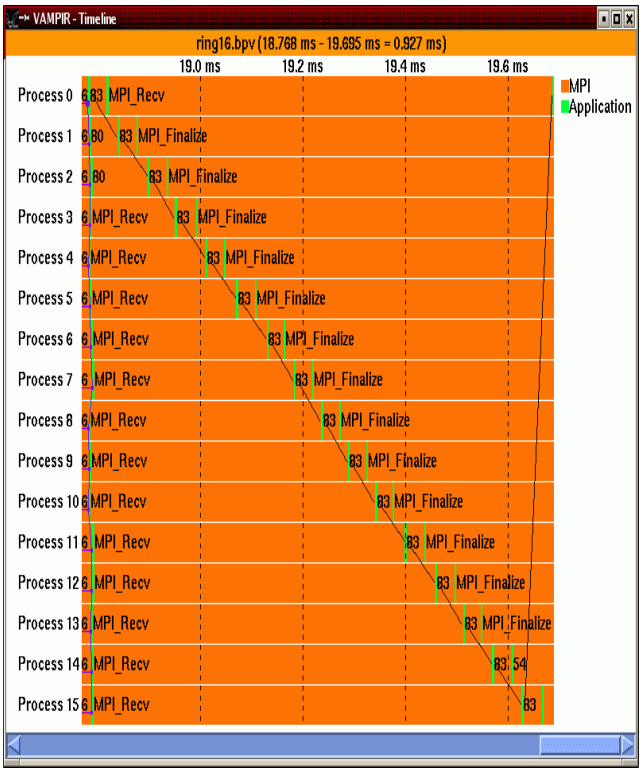
- Exascale systems will consist of
  - Complex configurations
  - With a huge number of components
    - Very likely heterogeneous

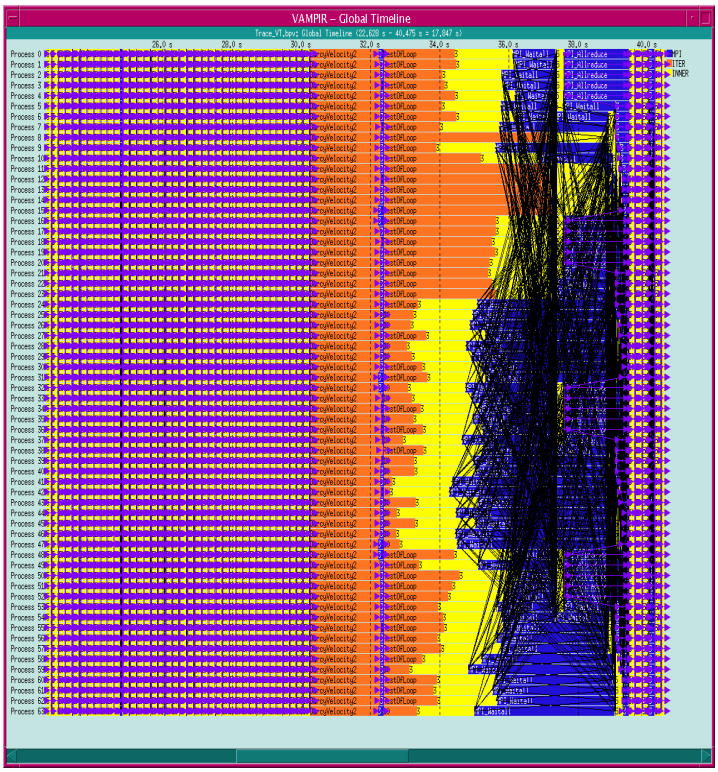
- Deep software hierarchies of large, complex software components will be required to make use of such systems

⇨Sophisticated integrated performance measurement, analysis, and optimization capabilities will be required to efficiently operate an Exascale system

⇨ Tools which provide **insight** not just numbers or charts needed!

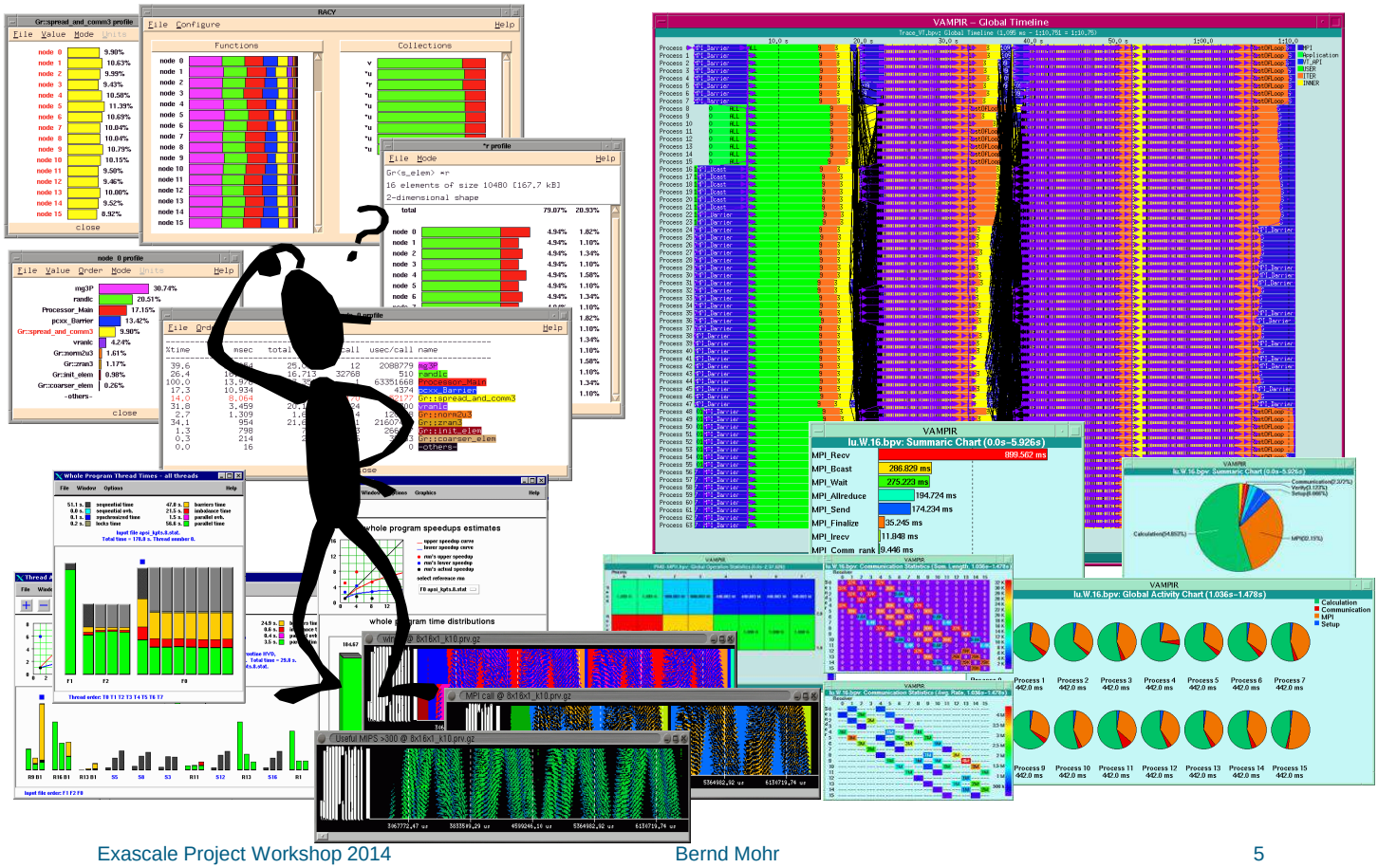
“A picture is worth 1000 words…”



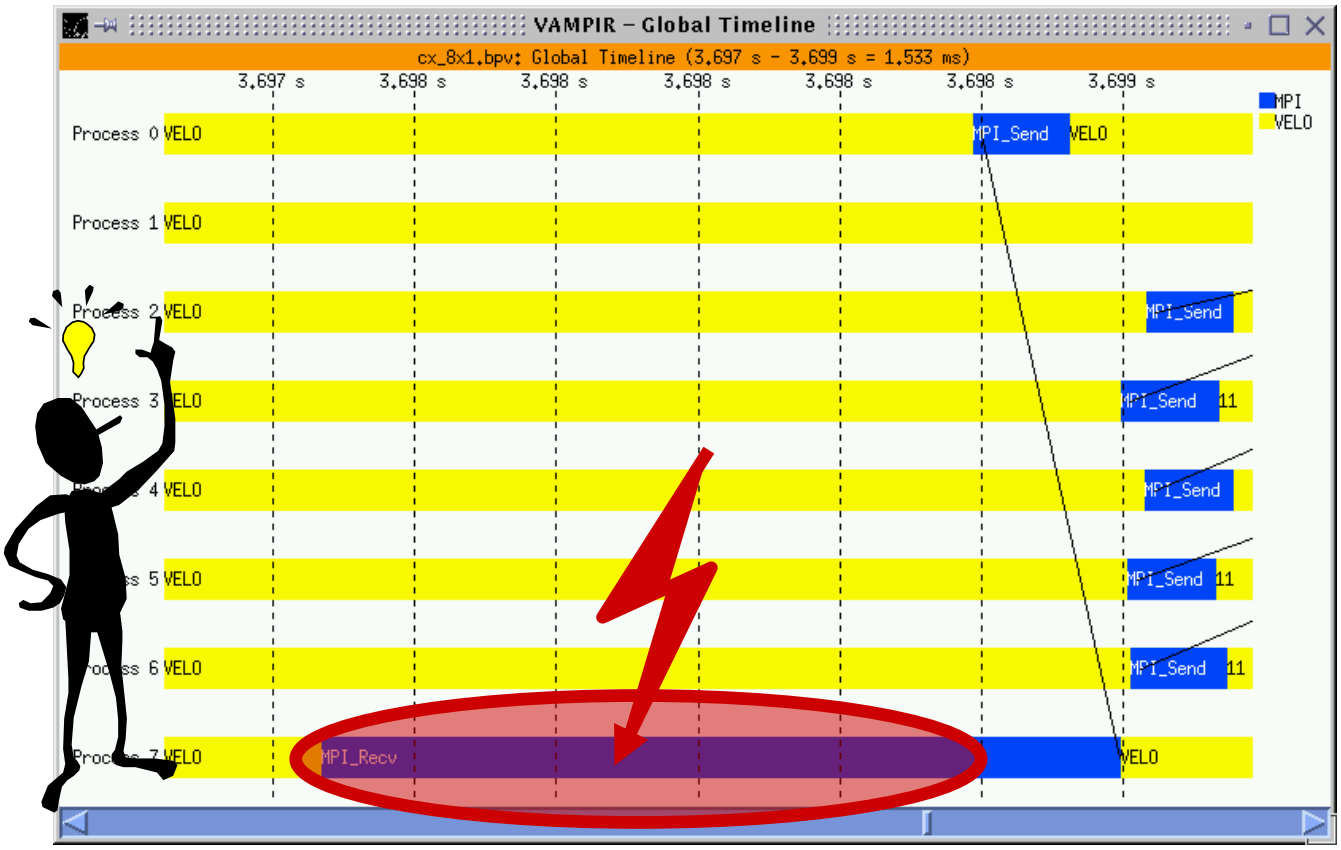


- MPI ring program
- “Real world” example

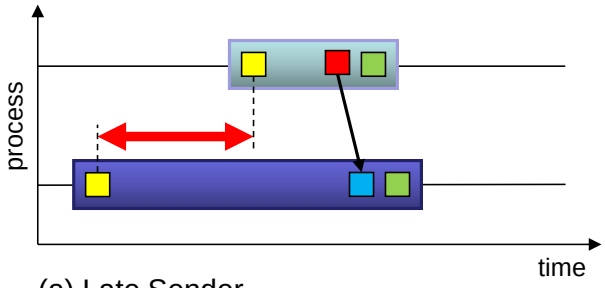
“What about 1000’s of pictures?” (with 100’s of menu options)



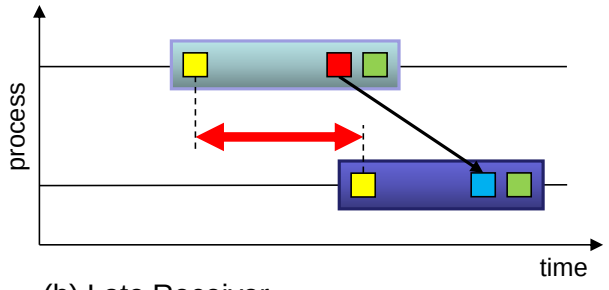
Example Automatic Analysis: Late Sender



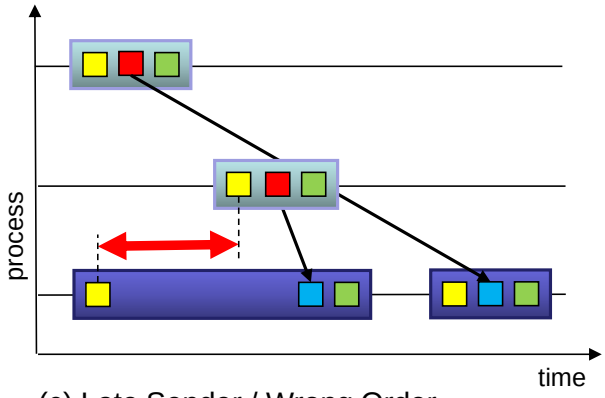
Scalasca: Example MPI Patterns



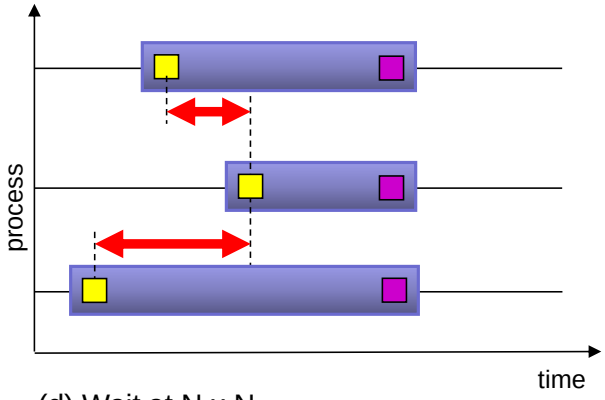
(a) Late Sender



(b) Late Receiver



(c) Late Sender / Wrong Order



(d) Wait at N x N

ENTER EXIT SEND RECV COLLEXIT

The Scalasca Project

JÜLICH FORSCHUNGSZENTRUM

scalasca

http://www.scalasca.org/

- Scalable Analysis of Large Scale Applications
- Approach
  - **Instrument** C, C++, and Fortran parallel applications
    - Based on MPI, OpenMP, SHMEM, or hybrid
  - **Option 1: scalable call-path profiling**
  - **Option 2: scalable event trace analysis**
    - **Collect** event traces
    - **Search** trace for event patterns representing inefficiencies
    - **Categorize and rank** inefficiencies found

JÜLICH FORSCHUNGSZENTRUM

NEW

Based on Score-P (CUBE4, OTF2)

German Research School for Simulation Sciences

Portability: Supported Platforms

JÜLICH FORSCHUNGSZENTRUM

- **Instrumentation and measurement only (visual analysis on front-end or workstation)**
  - Cray XT3/XT4/XT5, XE6, XK6
  - IBM BlueGene/L, BlueGene/P, BlueGene/Q
  - NEC SX8 and SX9
  - K Machine, Fujitsu FX10
  - Intel Xeon Phi
- **Full support (instrumentation, measurement, and automatic analysis)**
  - Linux IA32, IA64, x86_64, and PPC based clusters
  - IBM AIX Power3/4/5/6/7 based clusters
  - SGI Linux IA64 and x86_64 based clusters
  - SUN/Oracle Solaris Sparc and x86/x86_64 based clusters

Some known Installations of Scalasca

JÜLICH FORSCHUNGSZENTRUM

Companies
- Bull (France)
- Dassault (France)
- EDF (France)
- Efield (Sweden)
- GNS (Germany)
- MAGMA (Germany)
- RECOM (Germany)
- Shell (Netherlands)
- Sun Microsystems (USA)
- Qontix (UK)

Research / HPC Centres
- ANL (USA)
- BSC (Spain)
- CastorC (Cyprus)
- CEA (France)
- CERFACS (France)
- CINECA (Italy)
- CSC (Finland)
- CSCS (Switzerland)

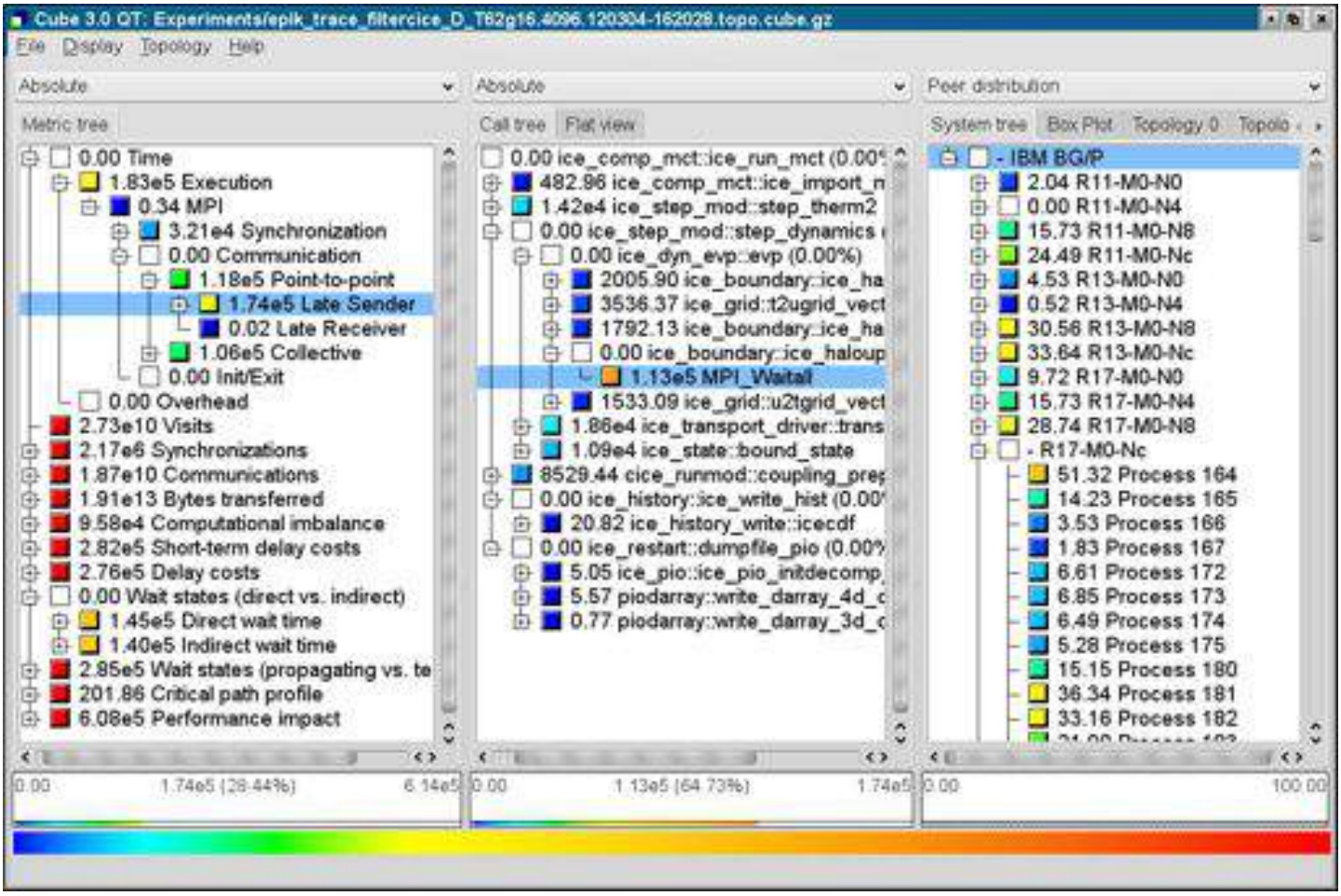
Research / HPC Centres (cont.)
- DLR (Germany)
- DKRZ (Germany)
- EPCC (UK)
- EPFL (Switzerland)
- HLRN (Germany)
- HLRS (Germany)
- ICHEC (Ireland)
- IDRIS (France)
- JSCC (Russia)
- LLNL (USA)
- LRZ (Germany)
- MSU (Russia)
- NCAR (USA)
- NCSA (USA)
- NSCC (China)
- ORNL (USA)
- PNNL (USA)
- PSC (USA)
- RZG MPG (Germany)

Research / HPC Centres (cont.)
- RIKEN (Japan)
- SARA (Netherlands)
- SAITC (Bulgaria)
- SNL (USA)
- TACC (USA)

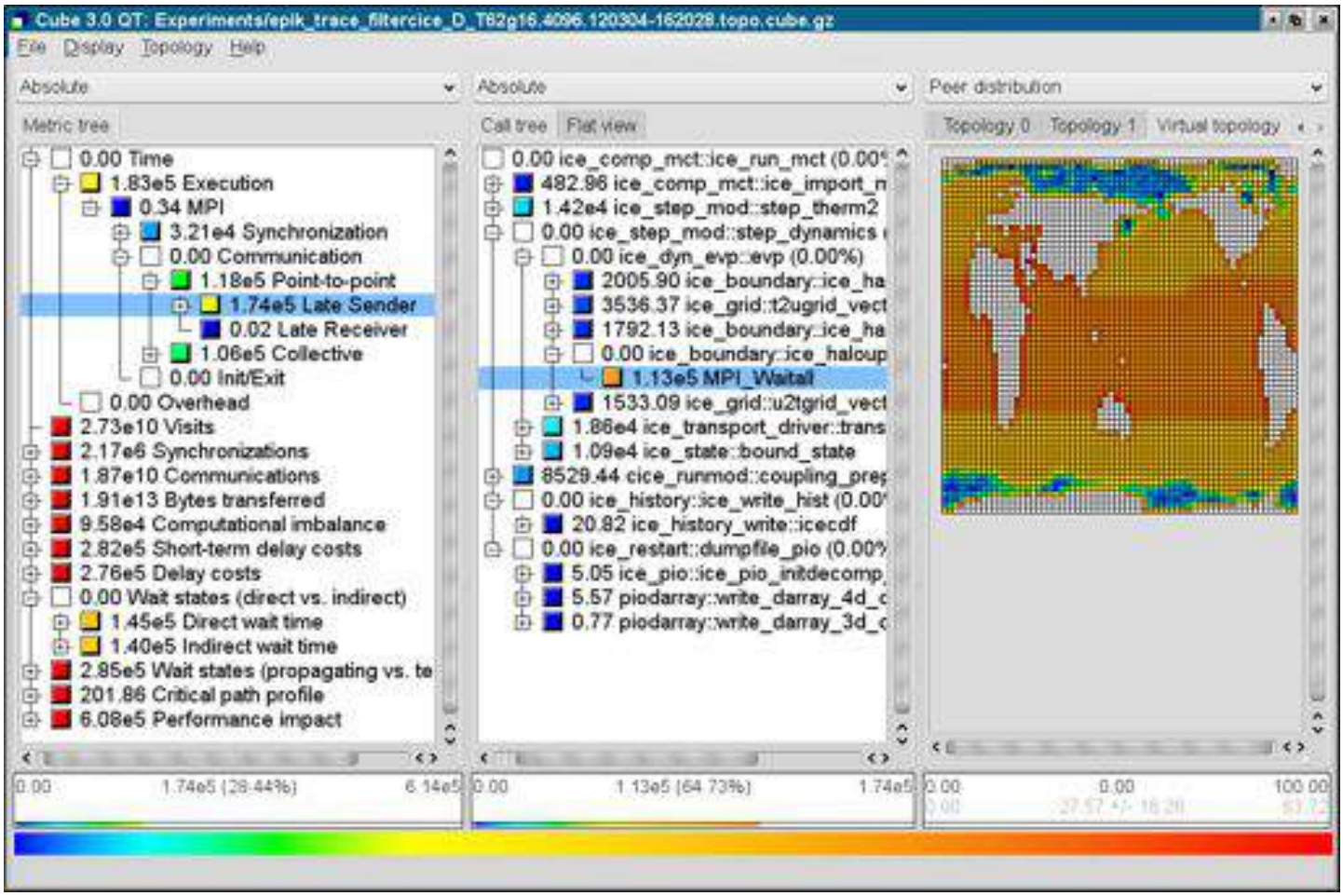
Universities
- GATech (USA)
- NUDT (China)
- RPI (USA)
- TITech (Japan)
- Germany Gauss Allianz Centres
- UOregon (USA)
- UTK (USA)
- **many more here …**

DoD Computing Centers (USA)
- AFRL, ARL, ARSC, ERDC, Navy, MHPCC, SSC-Pacific

Scalasca Example: CESM Sea Ice Module Late Sender Analysis

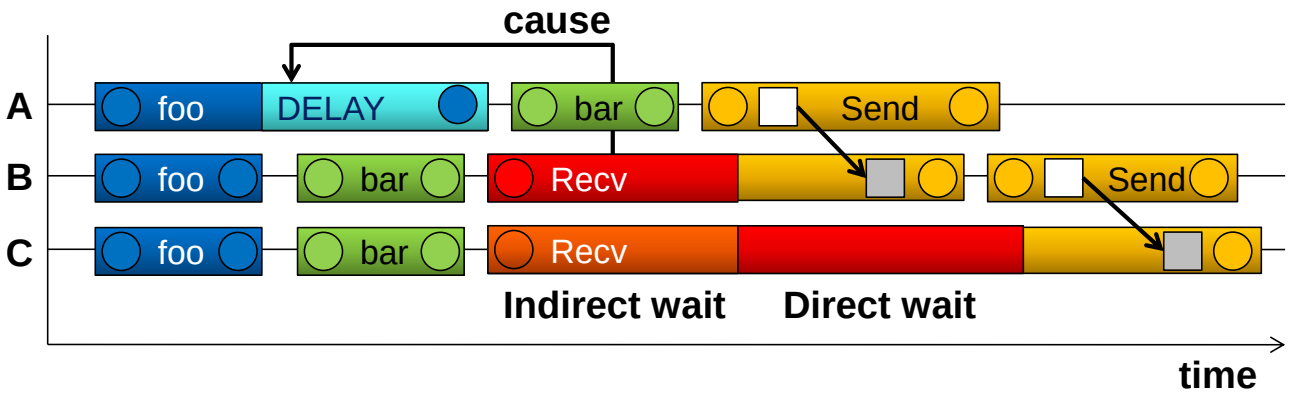


Scalasca Example: CESM Sea Ice Module Late Sender Analysis + Application Topology

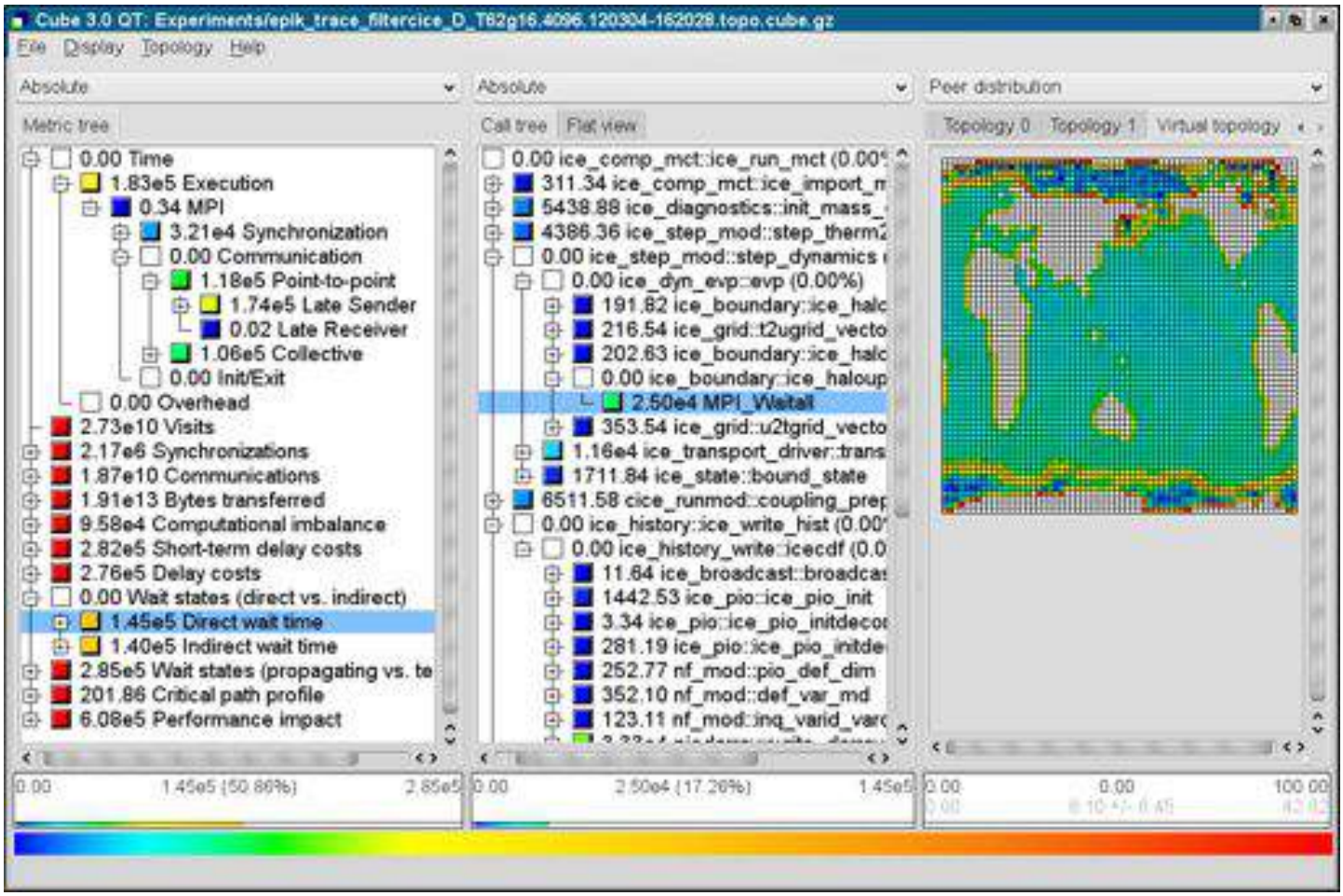


Scalasca Root Cause Analysis

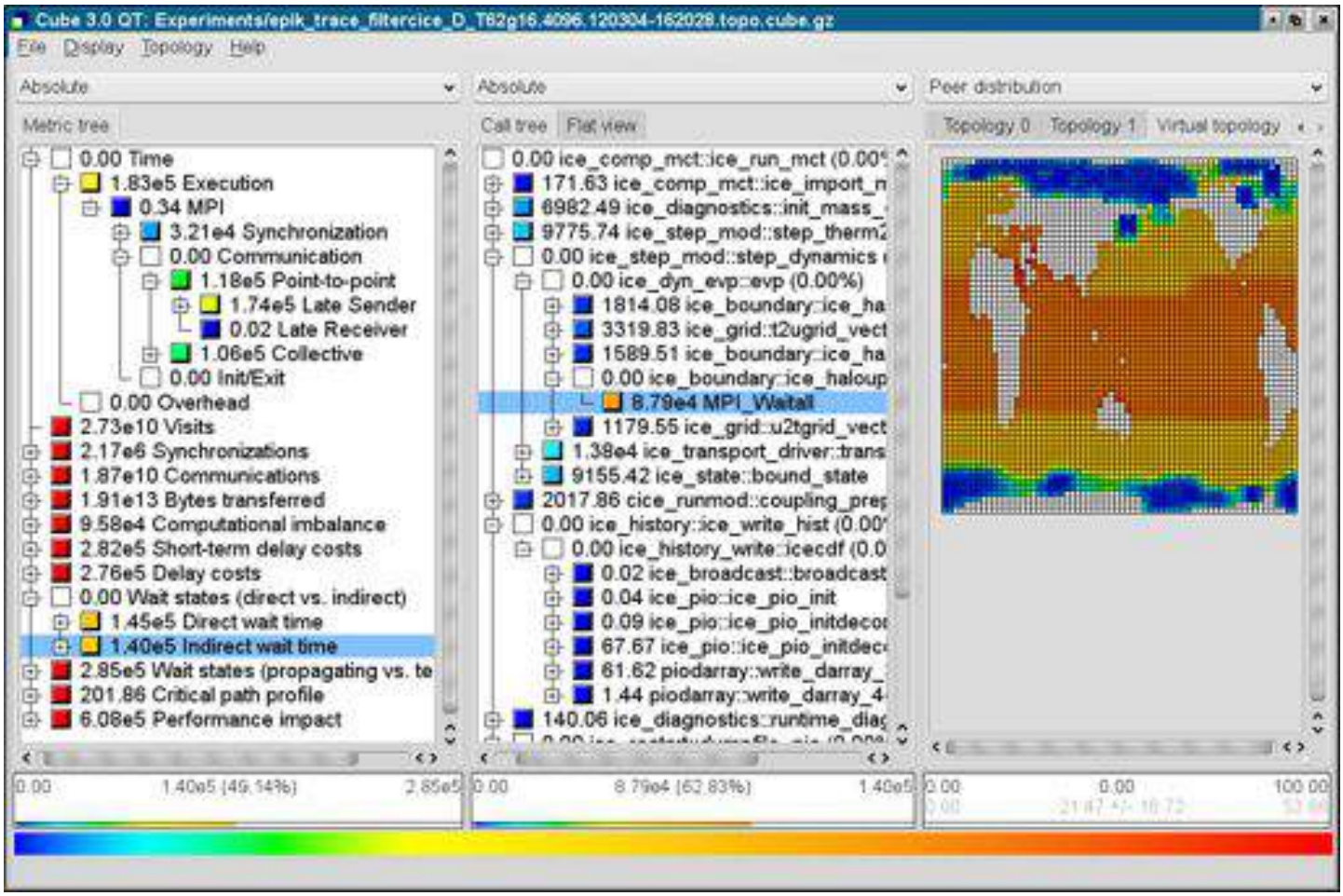
- **Root-cause analysis**
  - Wait states typically caused by load or communication imbalances earlier in the program
  - Waiting time can also propagate (e.g., indirect waiting time)
  - Goal: Enhance performance analysis to find the root cause of wait states
- **Approach**
  - Distinguish between direct and indirect waiting time
  - Identify call path/process combinations delaying other processes and causing first order waiting time
  - Identify original **delay**



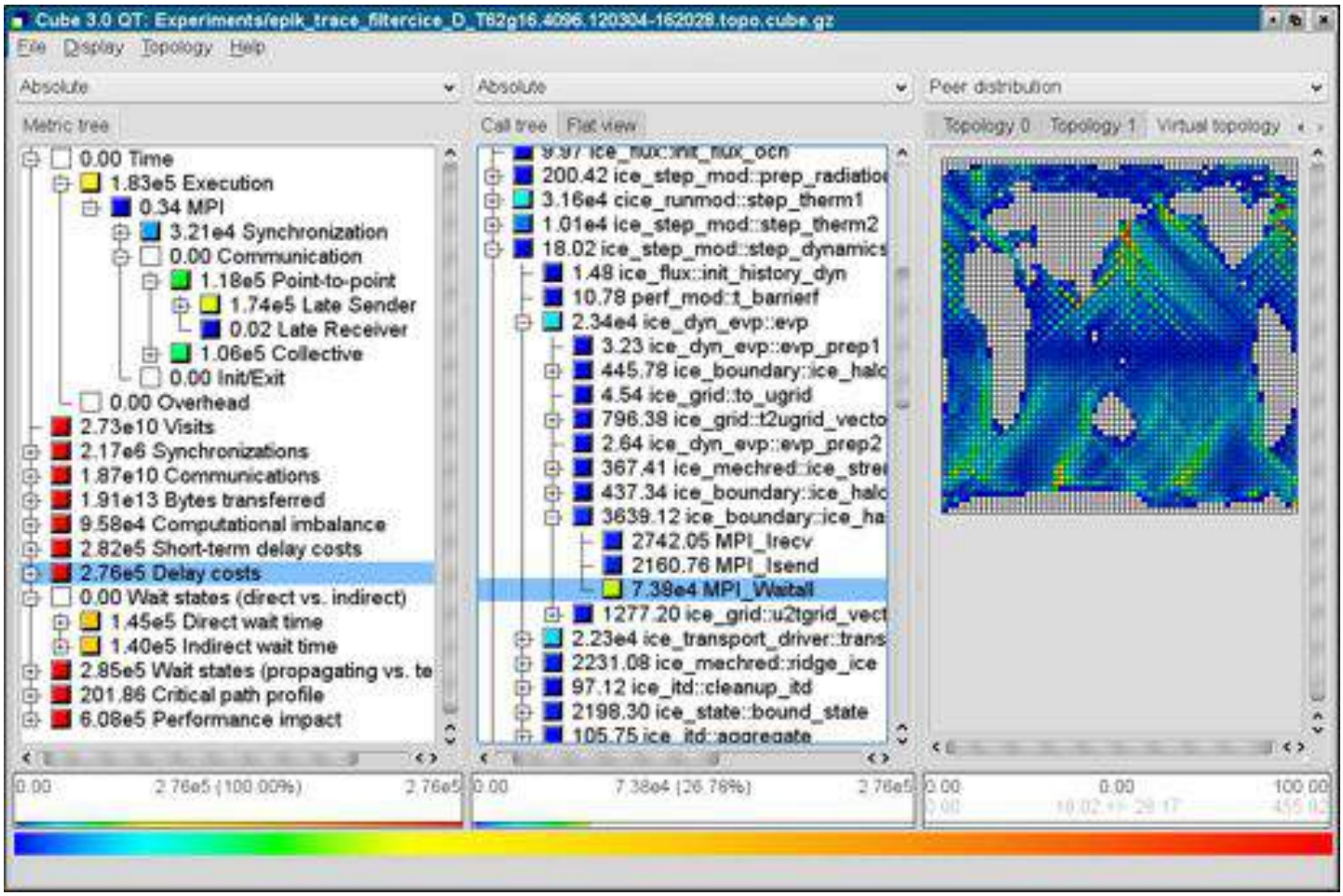
Scalasca Example: CESM Sea Ice Module Direct Wait Time Analysis



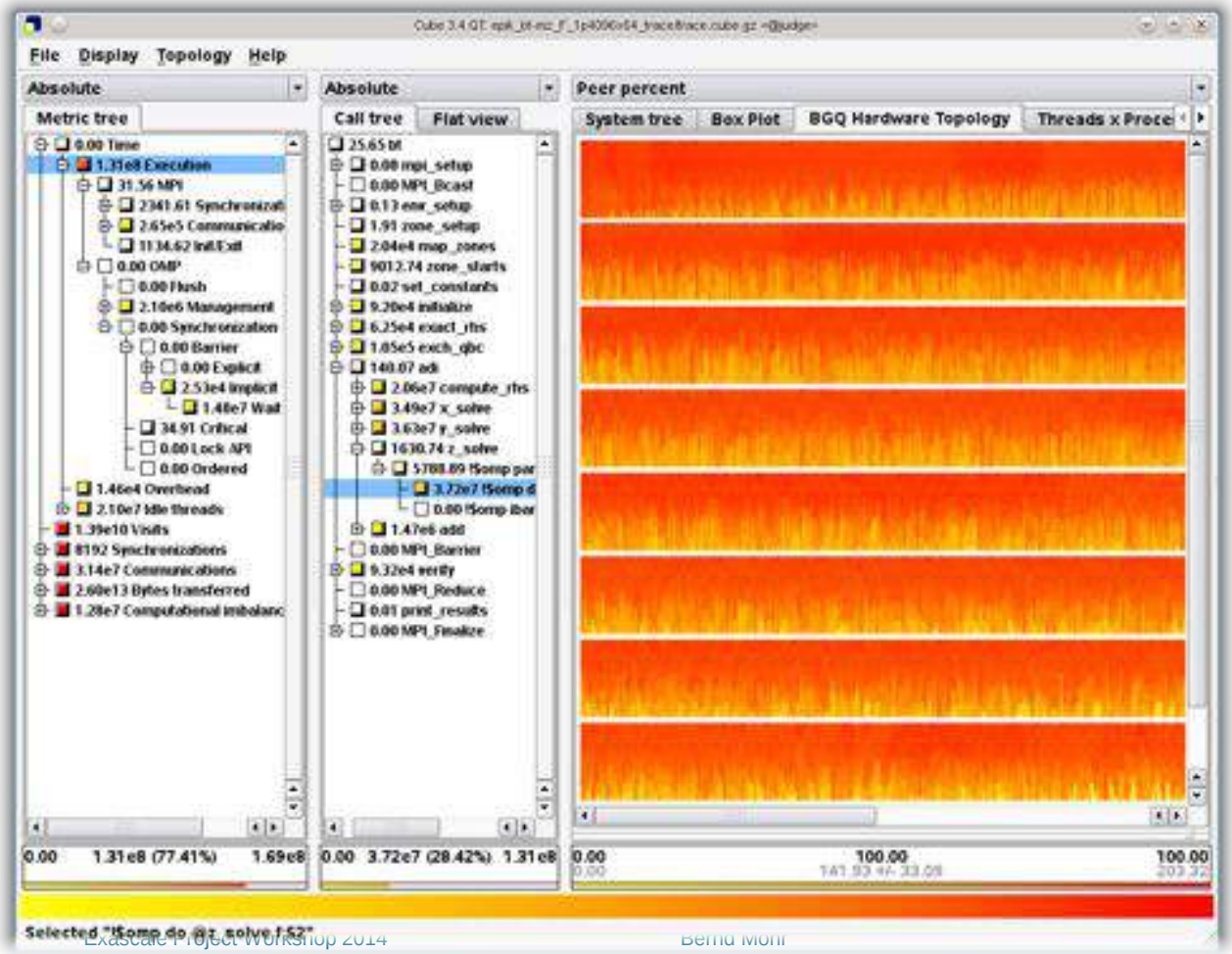
Scalasca Example: CESM Sea Ice Module Indirect Wait Time Analysis



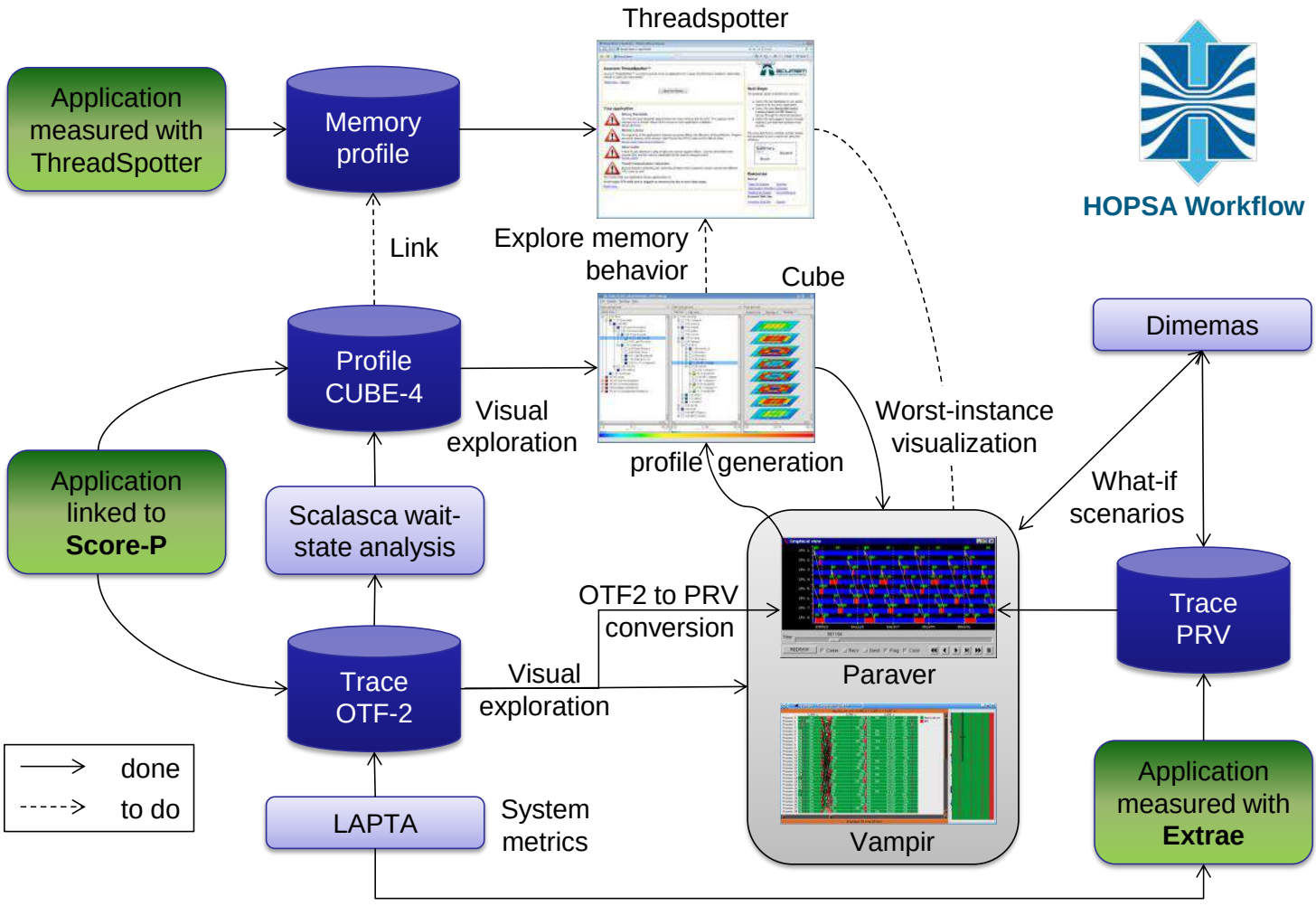
Scalasca Example: CESM Sea Ice Module Delay Costs Analysis



Scalasca trace analysis bt-mz@1,048,704 BGQ



Integration of Score-P based Tools



JÜLICH FORSCHUNGSZENTRUM

HOPSA Workflow

Projekt Work

JÜLICH FORSCHUNGSZENTRUM

- Intel Xeon Phi support
- Support for MPI_Spawn and Intercommunicators
- Support for Offload
- Support for OmpSs@DEEP

- Mont-Blanc1
  - Basic OmpSs support
- Mon-Blanc2
  - OmpSs/OpenCL + OpenCl support
  - Tighter integration with Paraver
  - OmpSs@cluster support

Acknowledgements

JÜLICH FORSCHUNGSZENTRUM

- **Scalasca team (JSC)**


Markus Geimer


Michael Knobloch


Bernd Mohr


Peter Philippen


Christian Rössel


Marc Schlütter


Pavel Saviankou


Alexandre Strube


Brian Wylie


Anke Visser


Ilja Zhukov

(GRS)


David Böhme


Alexandru Calotoiu


Marc-André Hermanns


Daniel Lorenz


Monika Lücke


Sergei Shudler


Felix Wolf

- **Sponsors**

HELMHOLTZ ASSOCIATION

ITEA 2 INFORMATION TECHNOLOGY FOR EUROPEAN ADVANCEMENT

SEVENTH FRAMEWORK PROGRAMME

Deutsche Forschungsgemeinschaft DFG

Bundesministerium für Bildung und Forschung

DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA

Questions?

JÜLICH
FORSCHUNGSZENTRUM

scalasca

- Check out **http://www.scalasca.org**
- Or contact us at **scalasca@fz-juelich.de**