Vision, Innovation, Network and Friendship

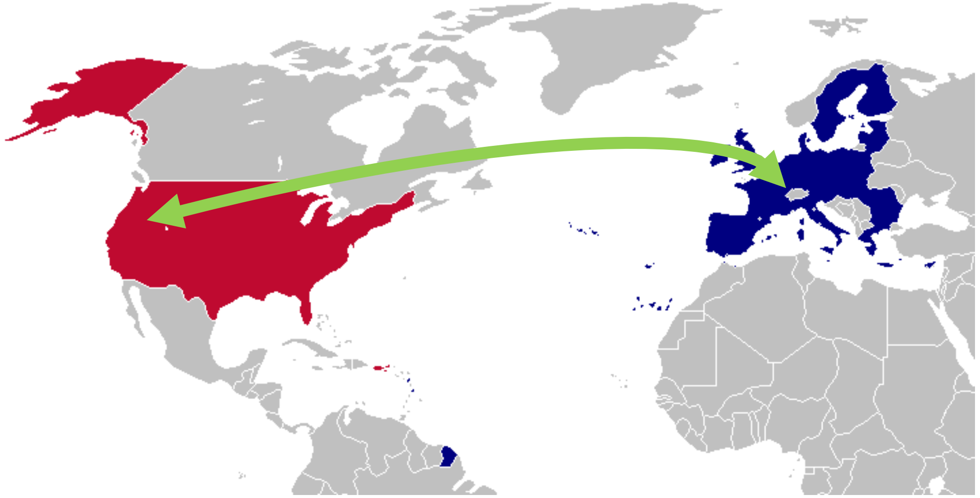
SOS20, Biltmore Inn, 25th March 2016

Marie-Christine Sawley

Marie-Christine.Sawley@intel.com

Celebrating SOS resilience

By one of the founders of the workshop series



EPFL-ETH Zurich Supercomputing scene

- 1986: first vector machine, CRAY 1s
  - 1989: Gigaflops award to CRPP code
- 1988: Cray 2@EPFL, CRAY XMP at ETH Zurich, 1st national strategy
- 1989: ETH Zurich-CSCS in Manno, national HPC machine
- 1992: Cray T3D, PATP collaboration
  - JPL
  - LLNL
  - PSC
- 1996: EPFL against T3E
- 1997: trip to Santa Fe
  - Ralf Gruber
  - Roberto Car
  - Michel Deville
  - Pierre Kuonen
  - Tony Gunzinger
  - Roland Richter
  - Marie-Christine Sawley

- Original traction
  - Plasma Physics
  - CFD
  - Material science
  - Big Science

Over the years,
SOS has proven to be a solid story of

✓Value

✓ Vision

✓ Innovation

✓ Network

✓

Friendship

The Swiss storyline

EPFL SUPERCOMPUTING REVIEW

EPFL-SCR No 9
Nov.97

CAPA

The SwissTx Supercomputer Project

Ralf Gruber, SIC-EPFL & Anton Gunzinger, IFE-ETHZ and SCS Zürich

Fig. 1 Santa Fe Workshop: Ken Kliewer (ORNL) organising US-Swiss working groups

En coopération avec certains groupes de recherche de l'EPFL et de l'ETHZ, nous proposons de développer, construire et d'installer les super-ordinateurs suisses Swiss-Tx qui sont entièrement basés sur des composants logiciel et matériel de série. La durée du projet s'étendrait jusqu'à la fin de 1999, date à laquelle il existerait une machine massivement parallèle atteignant une performance maximale

The Swiss storyline

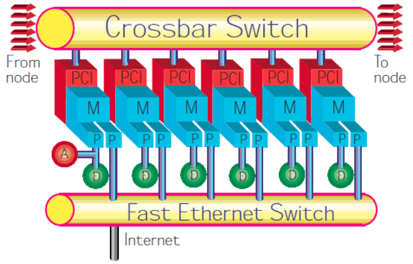


Fig. 1 – Swiss-T1 architecture is based on Alpha EV-6 processors. A node will consist of 6 dual processor boxes. They will be connected by a full 12x12 crossb`` switch based on the EasyNet concept. The remaining links are used to interconnect the 6 nodes. There will also be a Fast Ethernet. The users enter through the frontend which is fully integrated in the K-ring as a seventh node.

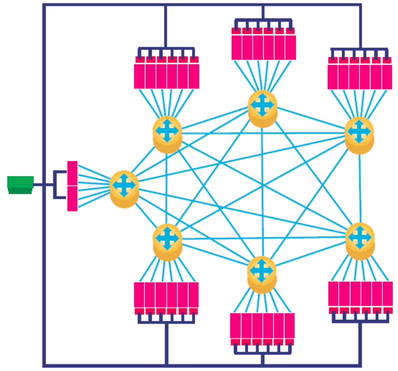


Fig.2 – The full connection between the 6 computational nodes and the integrated frontend at the left. The users will enter through the frontend computers

More integrated commodity supercomputers, that have a single-machine look, are presently developed at SRC, Sandia National Laboratories and EPFL. The SRC machine consists of a unique switch that interconnects Pentium Pros and later Merced processors. A first prototype will be installed at ORNL (Oak Ridge National Laboratories) beginning 99. The Sandia machine is based on 128 Alpha processors linked together with Myrinet switches. The EPFL machines are built in a cooperation with Supercomputing Systems in Zurich, Compaq/Digital, ETHZ and CSCS and are described below in detail. The SOS (Sandia/Oak Ridge/Swiss) research cooperation aims at shaping and testing the most promising supercomputer trends. In a previous paper (See EPFL Supercomputing Review n.10, 1997), we have presented the EPFL project, the Swiss-Tx. This communication gives the latest status of the systems development and the results of the first production runs for one scientific code, Speculoos.

The Swiss-Tx Commodity Supercomputer Project

The Partnerships

Highly evolved HPC relevant research projects have been conducted in Switzerland during the last 20 years. In hardware, the Institut für Elektronik (IFE, Prof. A. Gunzinger) at ETHZ and the Supercomputing Systems (SCS) company have developed the EasyNet concept, and have built two supercomputers, called Music and Gigabooster, both being commercialised through SCS. At EPFL, the PATP (Parallel Application Technology Program) project in cooperation with four major American research institutions and Cray Research had as goal the development

in implementation of the communication libraries (FCI, MPI-lite, MPI, and virtual shared memory programming model), in testing and evaluating the prototype machines (benchmarking), in porting/optimization of test programs in science, business and economy, in programming tools (monitors, debuggers, analysers), in a parallel file system and I/O, in distributed archiving, and in computational steering and visualisation. It is planned to commercialise the Swiss-Tx concept, in particular the crossbar switch that will be described later in more detail.

The Machines

Machine	T0	T0(Dual)	T1	T2
Date	Dec. 98	Sept. 98	1stQ99	1stQ00
#P	8	16	72	504
Peak Gflop/s	8	16	72	1008
Memory GBytes	2	8	36	252
Disk GBytes	64	170	800	5000
Archive TBytes	1	-	1	7
Operating system	DEC Unix	W NT	DEC Unix	not decided
Communi-cation system	EasyNet bus	EasyNet bus	12x12 crossbar	12x12 crossbar

And fame!

Haute école et supercomputer

L'Ecole d'ingénieurs du Valais se retrouve au cœur du «biocomputing». A Loèche-les-Bains la semaine passée, ses représentants ont dialogué avec les meilleurs mathématiciens du moment.

Avec les «boss» de la recherche de pointe en supercomputer: Al Geist, Bill Camp, Pierre Kuonen, Michael Levine, Neil Pundit.

Ils ont le contact avec les mathématiciens les plus évolués de la planète, en matière de recherches sur les superordinateurs: Pierre Kuonen de la Haute Ecole valaisanne et Ralf Gruber de l'EPFL.

A Loèche-les-Bains la semaine passée se déroulait la rencontre internationale «SOS». Cette abréviation réunit les instituts et les laboratoires les plus prestigieux des Etats-Unis dans le domaine de la recherche sur les superordinateurs. La lettre S est l'abréviation de Sandia national Labs, O l'abréviation de Oak Ridge National Laboratory et S veut dire Swiss. Et en Suisse, la pointe de la recherche se trouve à l'EPFL ou dans la Haute Ecole d'ingénieurs du Valais.

La rencontre fut organisée à l'initiative de celle-ci. Plus exactement de l'un de ses professeurs Pierre Kuonen.

Pierre Kuonen est un ancien de l'Ecole polytechnique fédérale de Lausanne. Il y a travaillé et enseigné à la grande époque des superordinateurs Cray, au milieu des années 1980. Il fut l'un des chercheurs de ce programme.

De cette époque, il a conservé des contacts étroits avec les meilleurs mathématiciens du moment dans le domaine: Bill Camp de Sandia, Al Geist d'Oak Ridge et Michael Levine du Pittsburg Supercomputing Center.

Avec son ami Ralf Gruber de l'EPFL, il a organisé à Loèche-les-Bains la rencontre internationale 2002. *«De fait, SOS se rencontre chaque année pour faire le point: soit outre-Atlantique, soit en Suisse»*, précisait Pierre Kuonen. *«Ces trois jours à Loèche-les-Bains sont un plus pour l'Ecole d'ingénieurs valaisanne.»* Surtout que les 35 invités sont tous des mathématiciens de pointe.

A l'époque des superordinateurs développés par l'EPFL en collaboration avec les laboratoires américains, leurs capacités étaient dévolues essentiellement à la physique et au nucléaire.

Depuis, la nouvelle frontière a changé. Elle s'appelle le «biocomputing». L'ambition actuelle est de suivre le transit des vitamines créées par l'ADN et l'ARN à travers les cellules du corps.

Cela demande des capacités de calculs bien plus grandes que celles offertes par les superordinateurs du style cray.

«De fait, il faut une puissance de calcul mille fois supérieure, soit 1 million de milliards d'instructions par seconde», précisait M. Kuonen.

Les termes scientifiques qui désignent ces gigantesques capacités sont imagés. Les supercomputers de l'ère Cray atteignaient une puissance de l'ordre du «Teraflop». Quant à la nouvelle génération, dévolue au biocomputing, elle atteindra une puissance de l'ordre du «Petaflop» (voir encadré).

Pascal Claivaz

Au cœur de la recherche actuelle

■ Avec les nouvelles recherches sur les superordinateurs, l'Ecole d'ingénieurs valaisanne et l'EPFL se trouve au centre des enjeux scientifiques et industriels actuels.

Les nouvelles puissances de calculs, du million de milliards d'instructions par seconde, serviront au développement des sciences de la vie. Or les biotechnologies connaissent un développement intense, depuis quelques années, en Valais et le long de l'arc lémanique.

De son côté, l'Ecole d'ingénieurs valaisanne tire profit de cette fréquentation avec les meilleurs esprits de la recherche en mathématiques de pointe.

Elle a notamment deux projets en cours. L'un concerne une analyse d'images en temps réels de patrons de tissus, pour en localiser les défauts et mettre en place les têtes de découpes.

L'autre consiste en une simulation de réseaux de télécommunications mobiles avec des partenaires comme Motorola, Telefonica España, l'université NTU d'Athènes, les Italiens d'ONNIS et INT France.

From SOS 1 until SOS12, topics say all

1997	• Santa Fe: *Build your own supercomputer*
1998	• Charleston*:*
1999	• Villars: *The Future of Supercomputers*
2000	• New Orleans
2001	• Hyannis Port: *Scalable Cluster Software*
2002	• Leukerbad: *Data Intensive Computing—health science*
2003	• Durango: *Architectural Considerations for Petaflops and Beyond*
2004	• Charleston: *Advanced Computer Architectures for Science*
2005	• Davos: *Full transition to MPP architectures*
2006	• Hawai*: Distributed and Green computing*
2007	• Key West*: High Throughput Computing*
2008	• Wildhaus

THOR'S HAMMER RED STORM

Excerpt from Bill Camp presentation, SOS8, Charleston

Issues in MPP Computing:

1. Physically shared memory does not scale
2. Data must be distributed
3. No single data layout may be optimal
4. The optimal data layout may change during the computation
5. Communications are expensive
6. The single control stream in SIMD computing makes it simple-- at the cost of severe loss in performance-- due to load balancing problems
7. In data parallel computing ('a la CM-5) there can be multiple control streams-- but with global synchronization

   Less simple but overhead remains an issue

8. In MIMD computing there are many control streams loosely synchronized (eg with messages)

   Powerful, flexible and complex

National Nuclear Security Administration

Sandia National Laboratories

Therefore……….is it the same story over again?

No

A number of important game changers!

Questions to address

- Are we entering a new age of software development for HPC?
  - Yes, since more than 25 years –to tell you how long I have been in this business!
  - Definitely an acceleration, and more roles/specialties →more funding for the “middleware”
- Application software longevity - a blessing or a curse?
  - Many newcomers come and go; bulk of HPC applications strongly rooted and evolving rapidly is the HPC Raison d’être
- What applications and workflows are driving HPC today?
  - HPC market revenue: BD, machine learning
  - HPC production: big science, engineering business
- Is co-design having an impact on system design?
  - yes, if it is understood that the pipe is long → no quick return
- How have HPC operating systems and runtime environments evolved?
  - Still room to grow

Game changers impacting 201X onwards

- Memory hierarchies
  - Workload, RT, OS, who is the driver?
- Application complexity; i.e. more attention paid to data structures
  - Application models are growing
  - Abstraction layers
- Workflows and usage models
  - Impact on designing and operating systems, policy makers
- HPC embraced by much larger community, with new workloads
  - Enhanced need to bridge with new specialties

What drives supercomputing market?

- In 2014, market update (source IDC)
  - HPC: 10 B$, 0.5 % of total IT market
  - Supercomputers, 3.2 B$, 0.16 % of total
  - Storage is the fastest growing segment of HPC, will continue with HPDA, according to IDC

The Broader HPC Market:
We Are Updating These Forecasts In December

- Now $11.4 Billion

The Broader HPC Market Growth to 2019

Worldwide HPC Compute, Storage, Middleware, Application and Service Revenues ($M)

	2014	2015	2016	2017	2018	2019	CAGR (14-19)
Server	10,222	~~10,718~~	11,467	12,958	14,073	15,165	8.2%
Storage	4,229	4,504	4,865	5,546	6,123	6,796	9.9%
Middleware	1,163	1,217	1,294	1,426	1,534	1,645	7.2%
Applications	3,598	3,769	4,028	4,479	4,824	5,167	7.5%
Service	1,819	1,895	2,006	2,223	2,356	2,497	6.5%
Total	21,032	22,103	23,660	26,632	28,910	31,270	8.3%

Source: IDC 2015

IDC Analyze the Future

Focus on application complexity

- Architectural features we can rely on for enhanced performance
  - Vectorisation (SIMD)
  - Instruction-level parallelism requires independent data sets within a loop
  - Pipelining is efficient on small regular loops
  - Branch prediction favour constant branch path
  - Prefetching (DRAM memory latency) favours contiguous stride -1 accesses
  - Caches favour data reuse, efficient if data structures allow
- Codes may exhibit on very brief time scale
  - Complex data dependencies (Stiff ODE solvers)
  - Dynamic data structures (AMR, multiresolution)
  - Data access patterns that hard to predict (HW)
  - Dynamic load imbalance
- And would not benefit from the features above
- Challenge
  - to identify regularity to expose the “right” granularity in order to benefit from such features
  - Phases where parallelism, computation demands, memory demands …, are “steady"

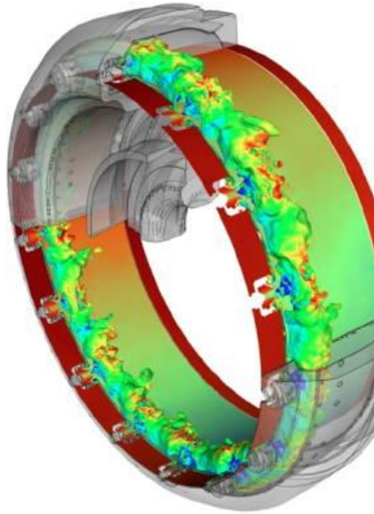
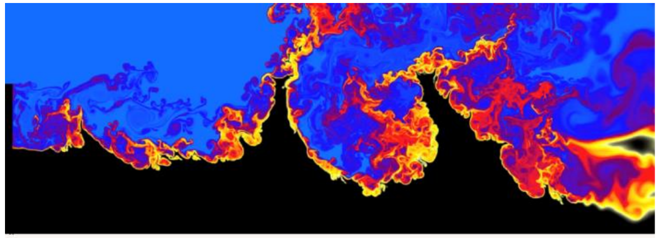


Image ref of very complex application: AVBP, CERFACS



Abstraction layers

- Growing number of collaborative studieson
  - Explore/propose ideas for abstraction mechanisms
    - **Isolate development** of new physics/algorithms from performance-sensitive operations
    - Allow **performance portability** across architectures,
  - Develop proof-of-concepts (PoCs) to test ideas for specific codes

- Abstraction/performance compromises
  - Abstraction which allows algorithmic optimizations? (re-using unused arrays for temp. storage, …) → memory copies?
  - Stay close to data structures (Fortran arrays, …)

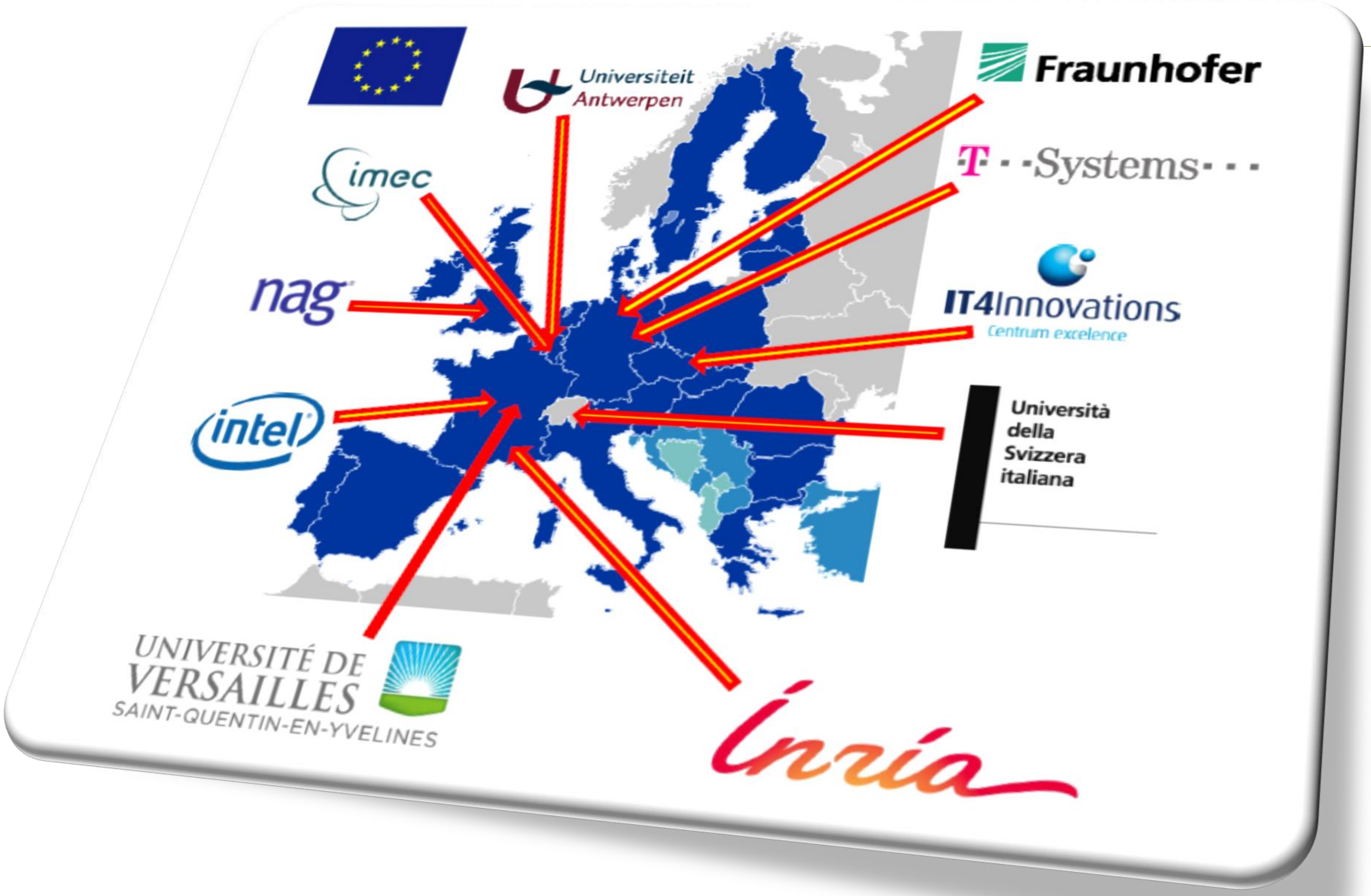
- Development choices
  - Programming language? (build system complexity, interfacing, adoption ….)
  - Abstraction without hindering physicist productivity?

- Stay pragmatic
  - Abstraction return on investment: decreases with abstraction level

Key message: code refactoring is very different than optimization

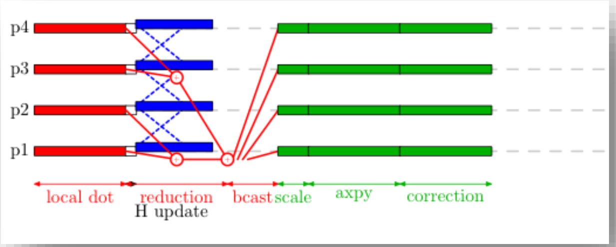
Example of joint effort in code refactoring

Partners



Universiteit Antwerpen
Fraunhofer
imec
T-Systems
nag
IT4Innovations
Centrum excelence
intel
Università della Svizzera italiana
UNIVERSITÉ DE VERSAILLES SAINT-QUENTIN-EN-YVELINES
Inria

EXA2CT



Solvers that scale to ExaScale

10^{18}

Programming models that scale to ExaScale

TBB

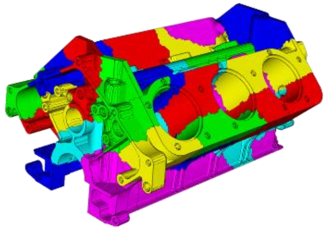
CILK

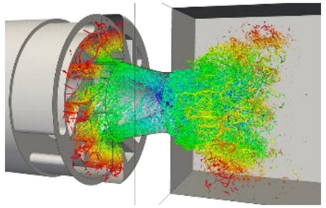
PATUS

GASPI

SHARK

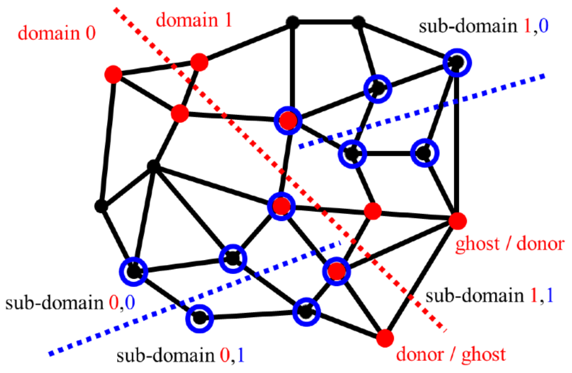
Using relevant real-life **proto applications**





Strong Exa-Scaling is Hard

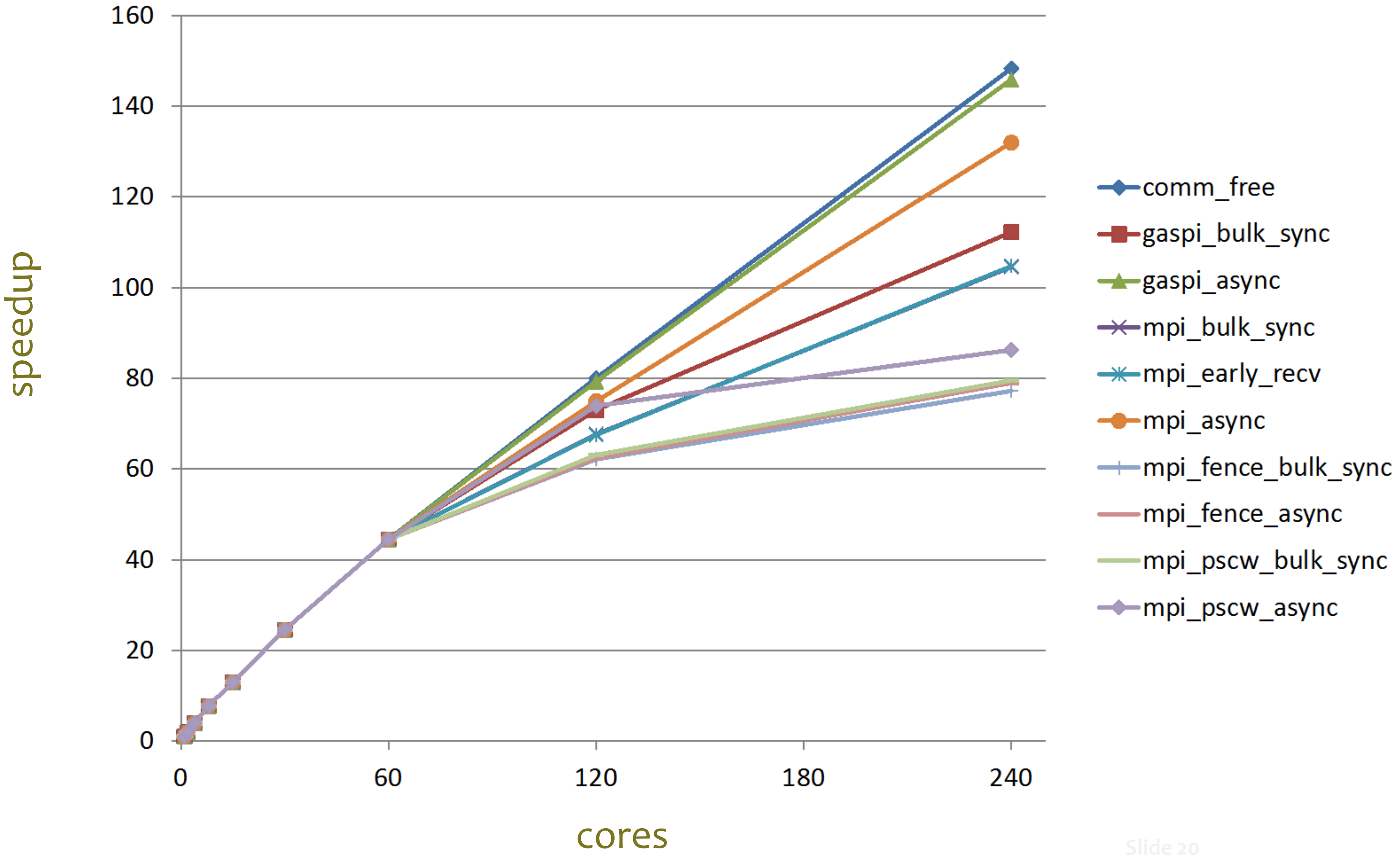
- CFD Application
  - Today: 50M mesh points
  - In ten years: 500M
- ExaScale Computers
  - 10M cores
  - Hence 50 mesh points per core
- CFD Proxy Application
  - Proto application of EXA2CT



T-Systems

DLR Deutsches Zentrum für Luft- und Raumfahrt e.V. in der Helmholtz-Gemeinschaft

CFD-Proxy on >1 Xeon-Phi



Proto Applications