DreamCloud: A new version of Supercomputing

…or life after the end of the Moore's Law…

Dr.-Ing. Alexey Cheptsov

OUTLINE

- **About us**
- **Convergence of Supercomputing into Big Data**
- **Semantic Web as a new HPC domain?**
- **Mission Data-Centric Parallel Programming Models**
- **DreamCloud project approach**

1 About HLRS

HLRS

Baden-Württemberg

KÄRCHER

Turnover: ~1-2 bil. €
Staff: ~8000

BOSCH

Turnover: ~50 bil. €
Staff: ~300.000

DAIMLER

Turnover: ~100 bil. €
Staff: 260.100

MAHLE

Driven by performance

Turnover: ~5 bil. €
Staff: ~50.000

STIHL

Turnover: ~2,5 bil. €
Staff: ~11.00

PORSCHE

Turnover: ~14 bil. €
Staff: ~19.000

1 About HLRS

High Performance Computing Center Stuttgart

- First Cray system in Europe
  (Cray-2, 1986, 4 CPUs, 2GB RAM, approx. 2 GFLOPS)
- National HPC infrastructure provider since 1995
- EU infrastructure provider since 2005
- **110M core hours delivered to industry in 2014**

HORNET (Cray XC40, Intel Haswell CPU, Aries network)

- **4.000** nodes (24 cores)
- **4 PFLOPs** performance
- **128 GB** RAM per node
- **7,8 PB** Disc
- **1512 KW** power consumption / 1.5M Euro

1 About HLRS

TOP500

Total:

- USA – 233
- Japan - 39
- Germany – 37
- China - 37

Newcomers in 2015:

- USA – 34
- Germany – 12
- Japan - 11

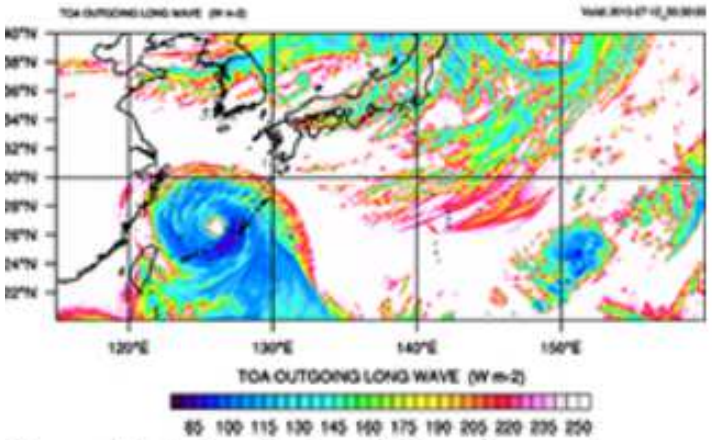
top500.org/lists/2015/06/ Search

RANK	SITE	SYSTEM	CORES	RMAX (TFLOP/S)	RPEAK (TFLOP/S)	POWER (KW)
1	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945

1 About HLRS

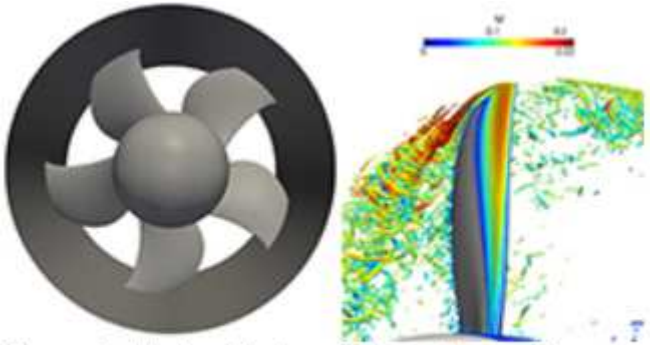
XXL Application Portfolio

- Climate simulation on a high (a few kms) resolution
  - University of Hohenheim
  - 84.000 cores, 84 hours, 450 TB data



Copyright: *Institute of Physics and Meteorology, Universität Hohenheim*

- Turbulent flow simulation for air- and gas dynamic analysis
  - RWTH Aachen University
  - 92.000 cores, 110 hours, 80 TB data



Copyright: *Institute of Aerodynamics, RWTH Aachen University*

See more at:
http://www.gauss-centre.eu/gauss-centre/EN/Projects/XXL_Projects_Hornet/XXL_Projects_Hornet.html?nn=1236240

Maya, the bee (みつばちマーヤの冒険)

- 115.000 pictures
- 2 hours each
- 200 pictures in parallel
- From 9600 days to 48 days
- 400.000 viewers in 2 weeks

Challenges for New-Generation Systems

- Computation power
  - Exascale is on his way, perhaps by 2020
  - More performance for less money
  - Hazelhehn - approx. **8 PTFLOPs** / **3M Euro** power costs, approx. 1 PB RAM
  - Sustainable performance - **1 PTFLOP**
- Storage
  - Spinning devices will go away by 2020
  - Flash is the core technology (aka NVM)
  - Tapes will remain for data persistency
- **Memory**
  - both high memory capacity and high memory bandwidth are required
  - cannot guarantee same growth as for FLOPS
  - extremely deep hierarchy (at the cache level)
  - programming ease
  - 3D stacked memory, NVM

Technology	latency	slow down
DRAM	20 - 50 Nanoseconds	1X
NVM (MRAM, other new technologies)	5 - 3000 nanoseconds	1/4X - 60X
SSD (NAND flash)	20,000 - 40,000 nanoseconds	1000X - 8000X
Magnetic disk	3,000,000 - 6,000,000 nanoseconds	150,000X - 1,200,000X

Source:

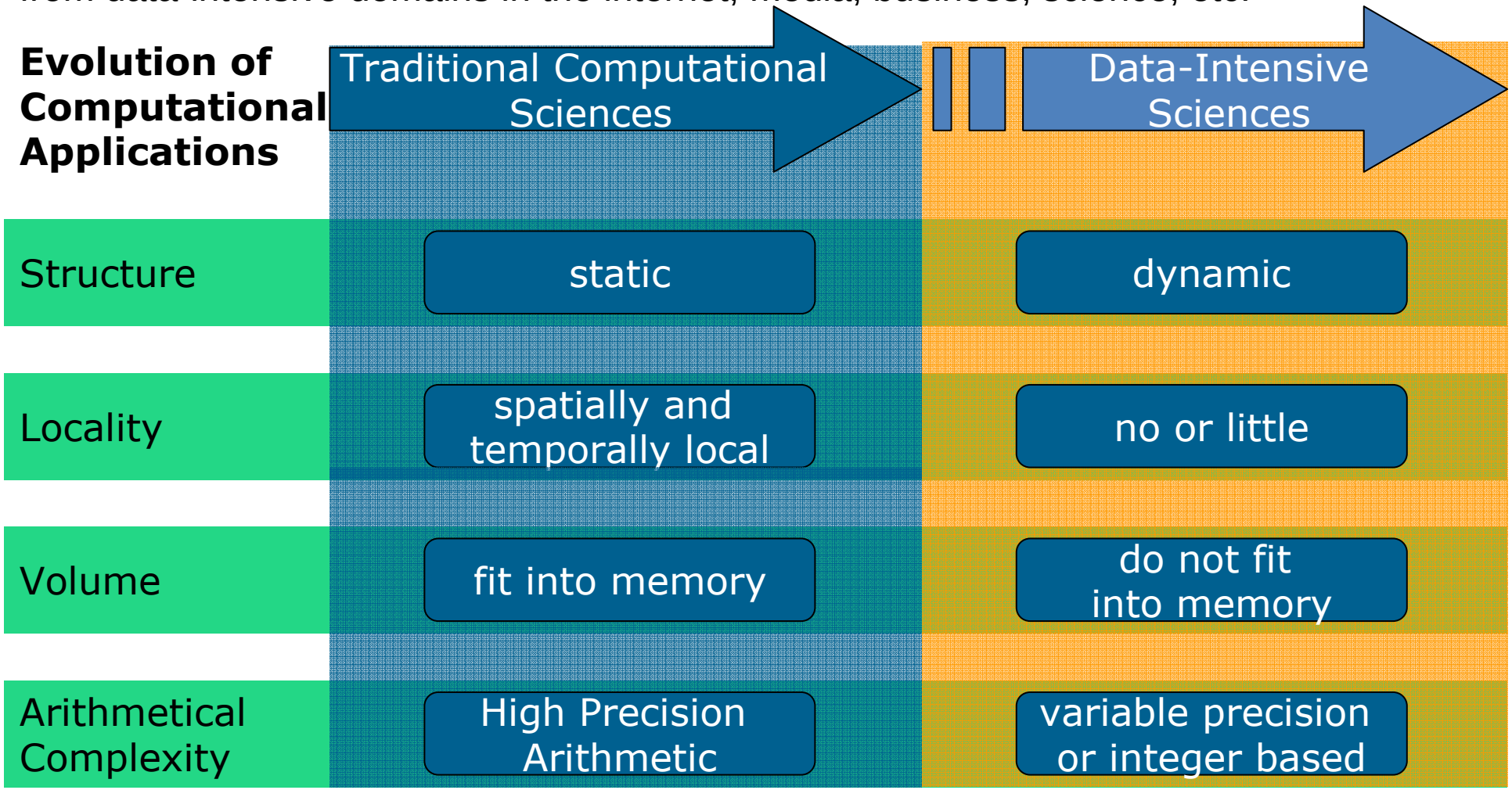
http://insidehpc.com/2014/05/nvm-will-shake-supercomputing/

OUTLINE

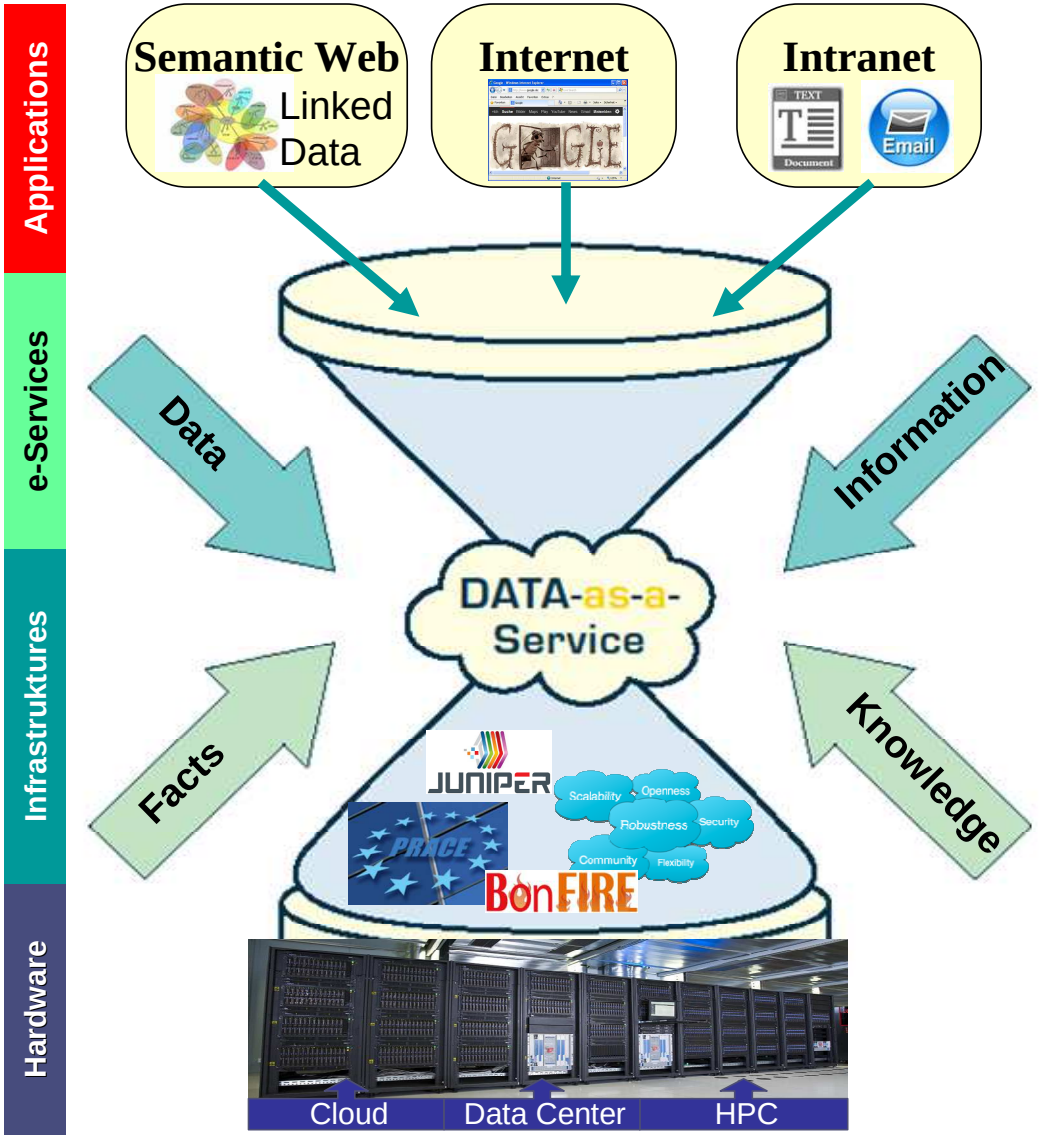
- **About us**
- **Convergence of Supercomputing into Big Data**
- **Semantic Web as a new HPC domain?**
- **Mission Data-Centric Parallel Programming Models**
- **DreamCloud project approach**

2 Convergence of Big Data into Supercomputing

The modern HPC have to address a new class of computing-intensive applications from data-intensive domains in the internet, media, business, science, etc.



2 Convergence of Big Data into Supercomputing



OUTLINE

- **About us**
- **Convergence of Supercomputing into Big Data**
- **Semantic Web as a new HPC domain?**
- **Mission Data-Centric Parallel Programming Models**
- **DreamCloud project approach**

3 Semantic Web as a New HPC Domain ?

H L R S

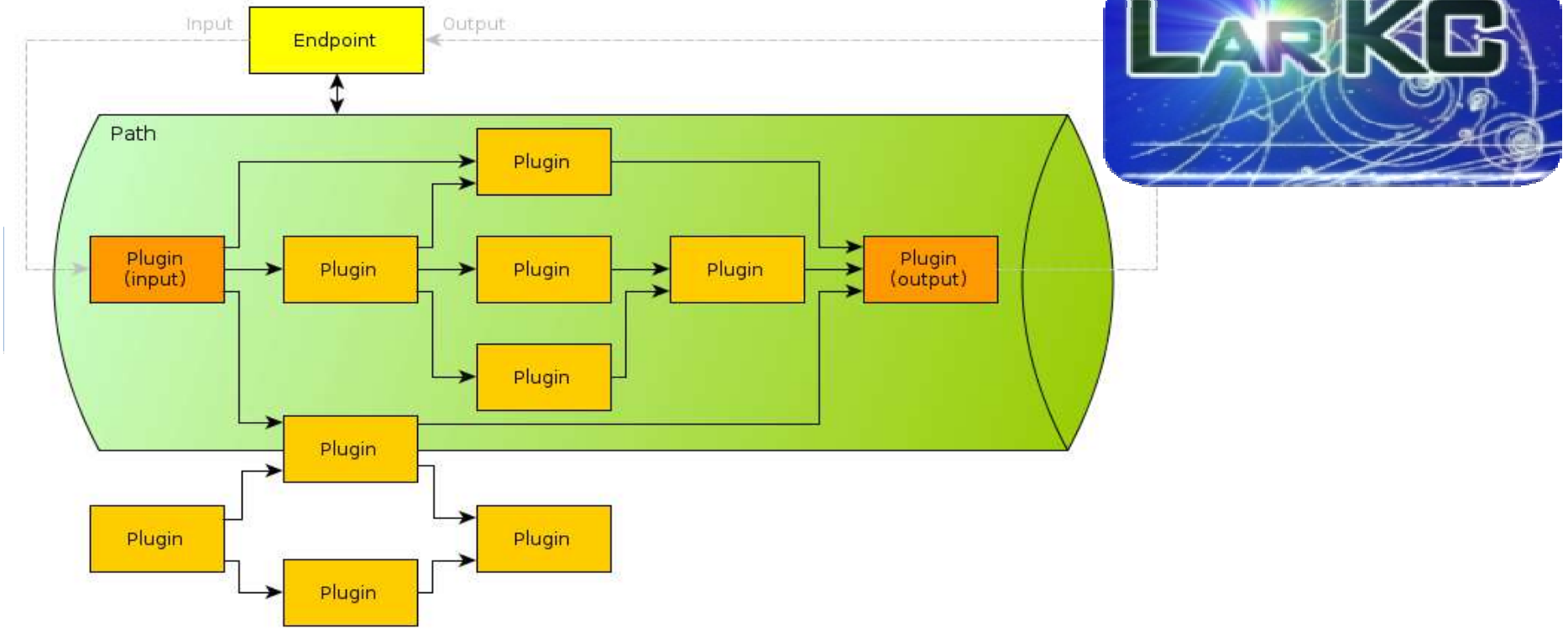
What are the challenges

- Infrastructure „on demand“
  - distributed memory parallel clusters with low-latency intercon.
  - multicore machines with shared memory
  - GPGPU devices
  - alltogether?

3 Semantic Web as a New HPC Domain ?

What are the challenges

A Semantic Web Integration
Platform for Large-Scale Reasoning



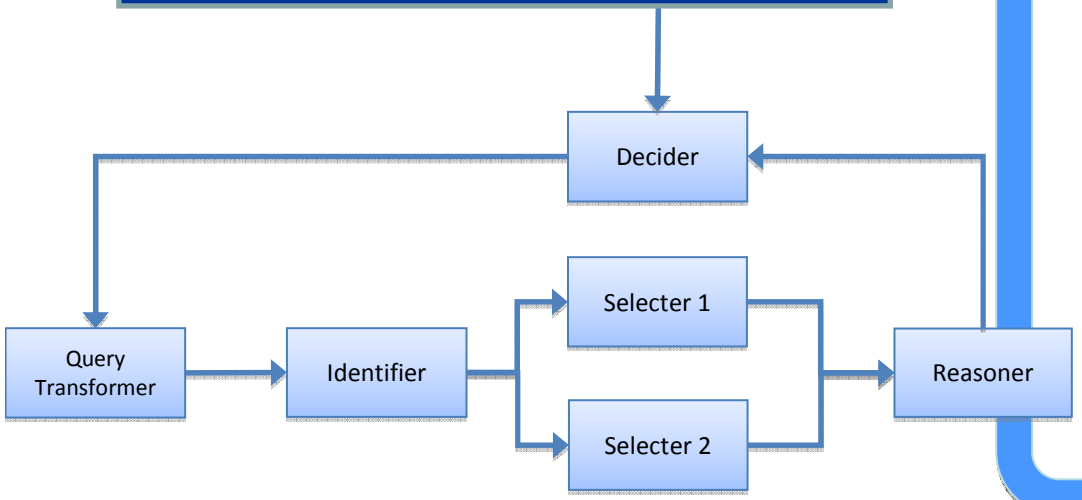
3 Semantic Web as a New HPC Domain ?

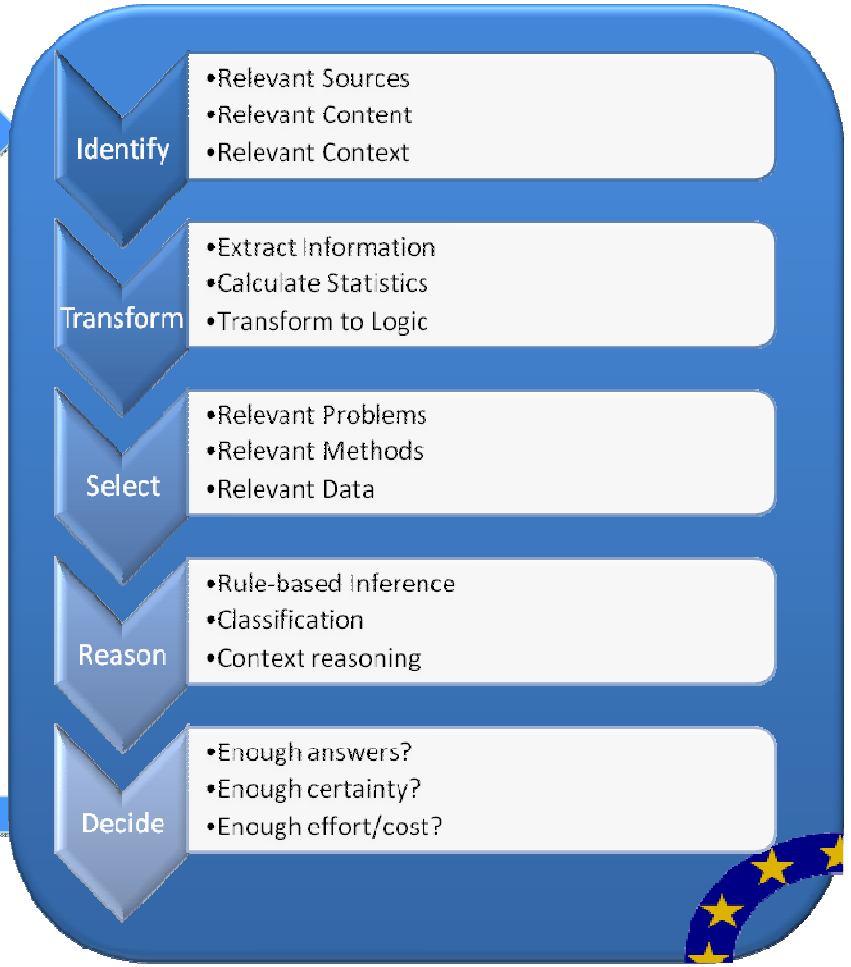
H L R S

Development Platforms

➢ The idea of LarKC

LarKC = an infrastructure for **large scale, high performance** incomplete **reasoning**





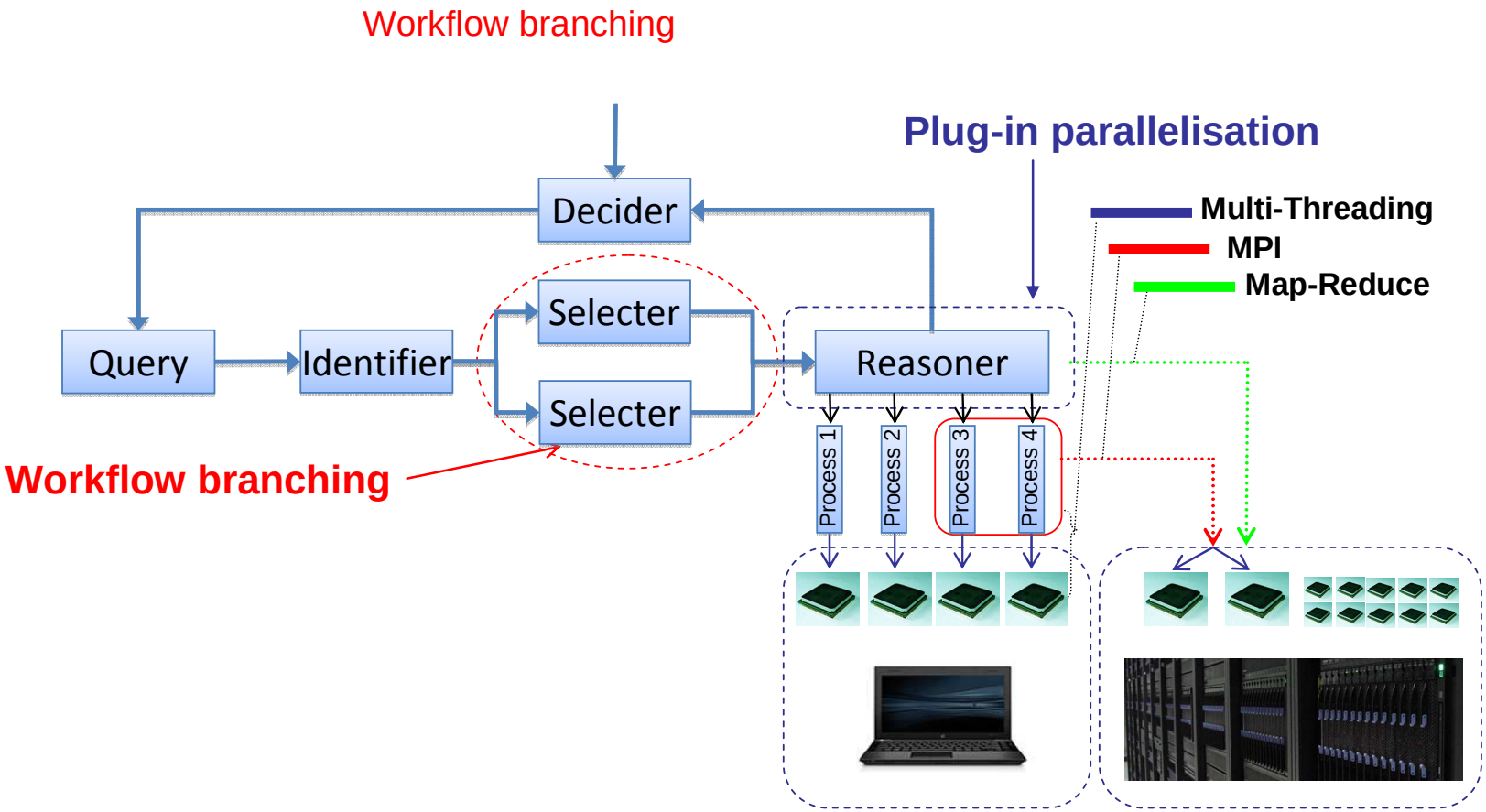
Flexibility, Modularity
Scalability, Performance

3 **Semantic Web as a New HPC Domain ?**

H L R S

Development Platforms

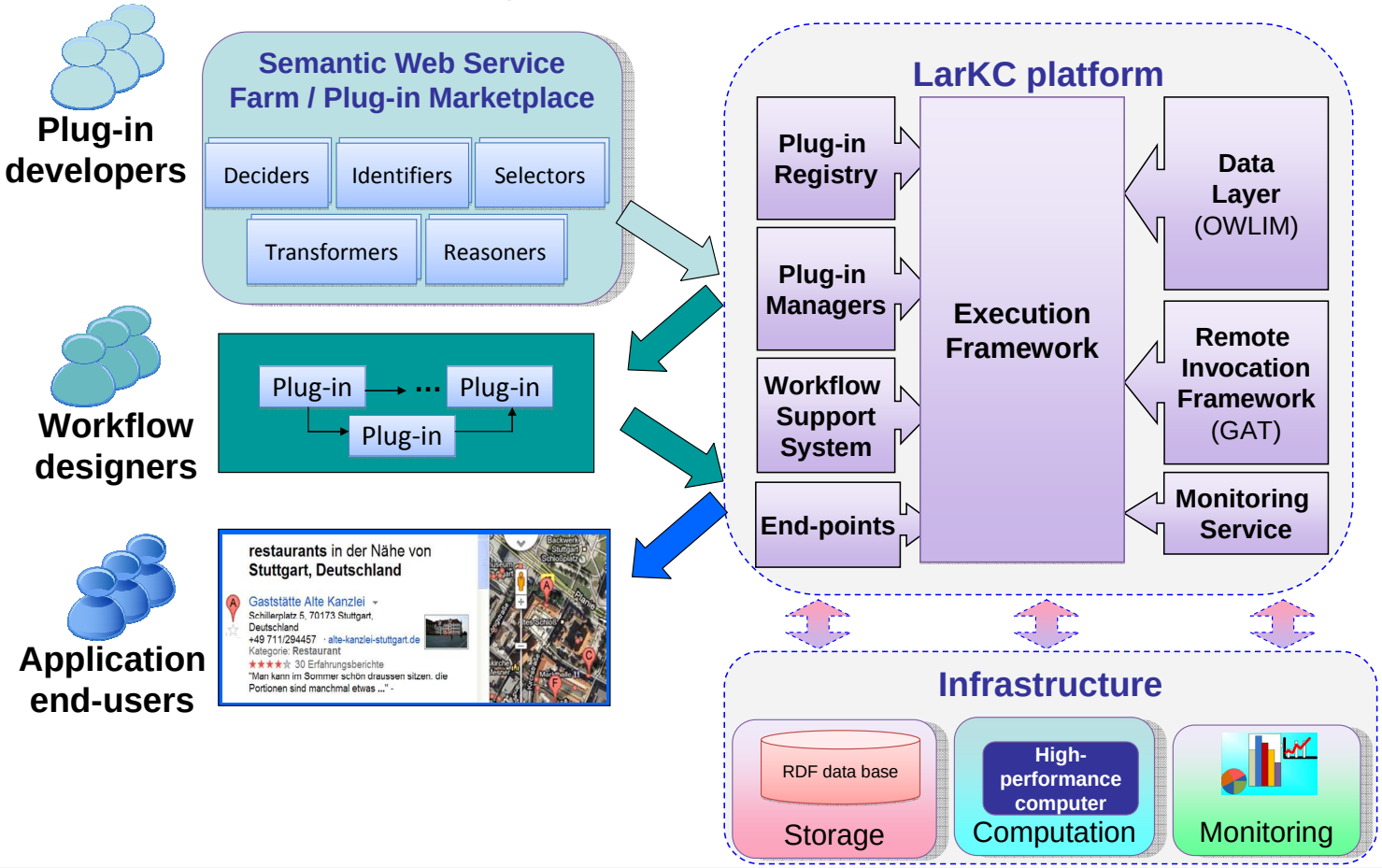
➢ LarKC architecture: High-Level Overview



3 Semantic Web as a New HPC Domain ?

Development Platforms

➢ LarKC architecture: High-Level Overview



OUTLINE

- **About us**
- **Convergence of Supercomputing into Big Data**
- **Semantic Web as a new HPC domain?**
- **Mission Data-Centric Parallel Programming Models**
- **DreamCloud project approach**

State-Of-The-Art: Distributed Memory Processing over FS

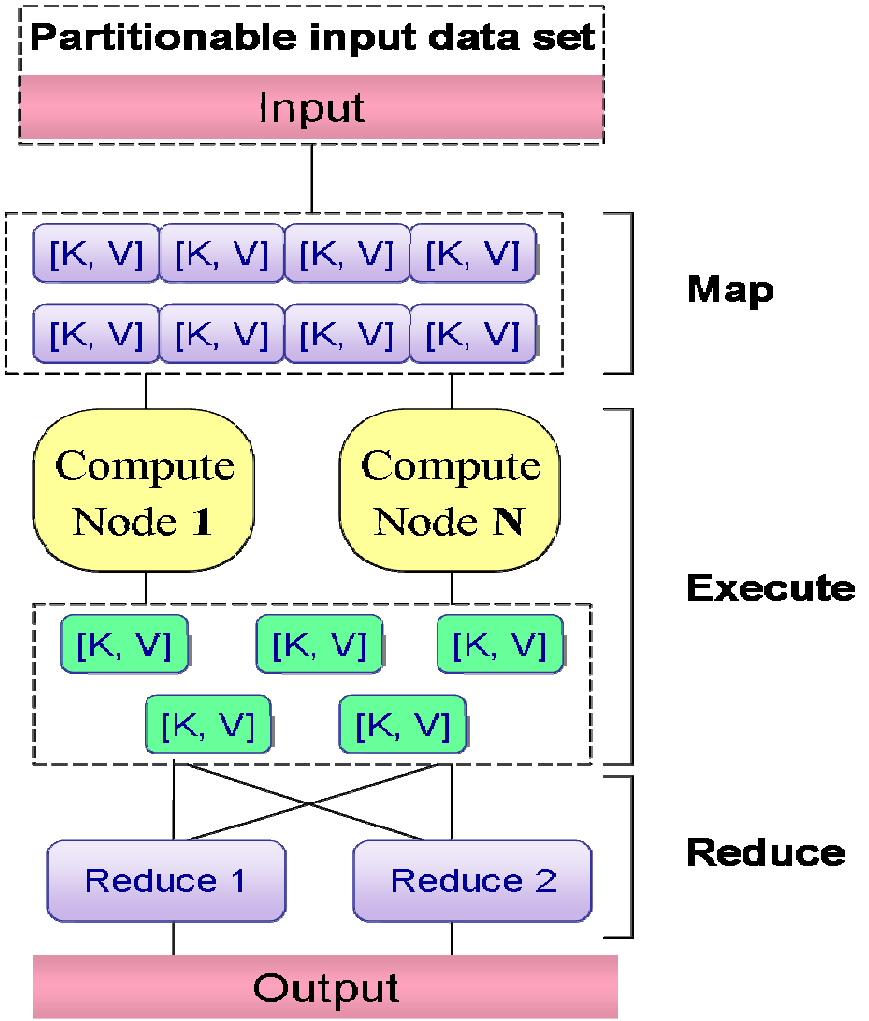
➢ Programming models:
MapReduce


data-centric


fault-tolerant


poor sustainable performance


restrictive key/value model

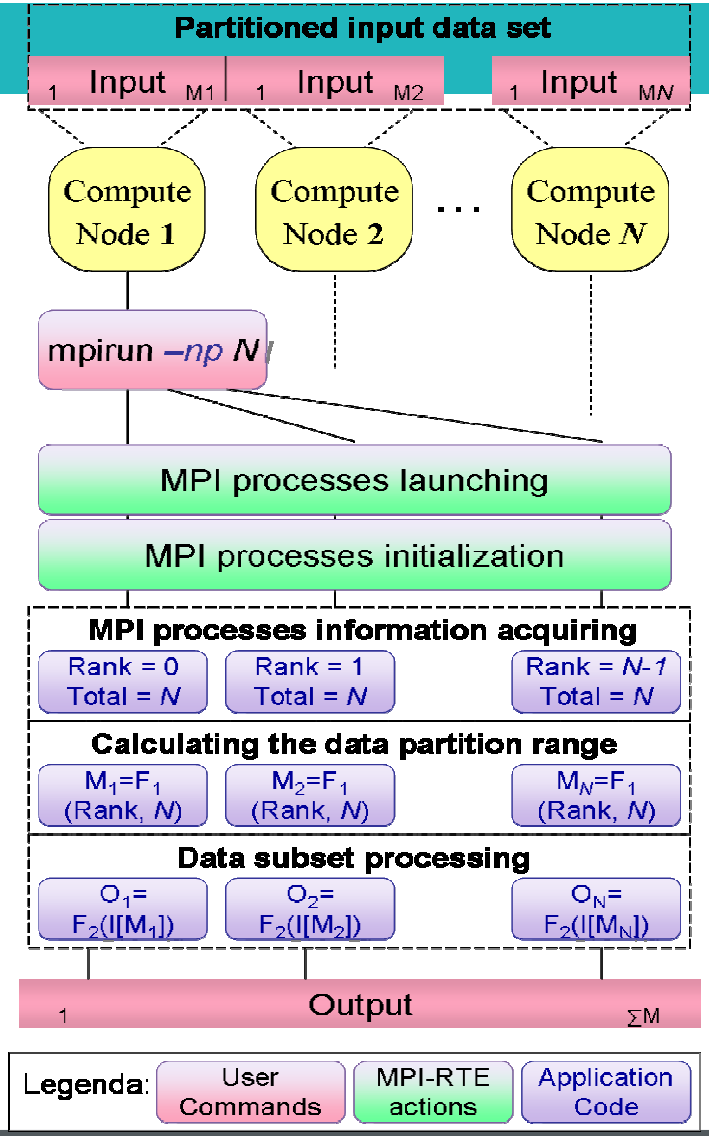


The Message-Passing Interface

- Data-driven scenarios with MPI

easy to integrate

high performance



The Message-Passing Interface

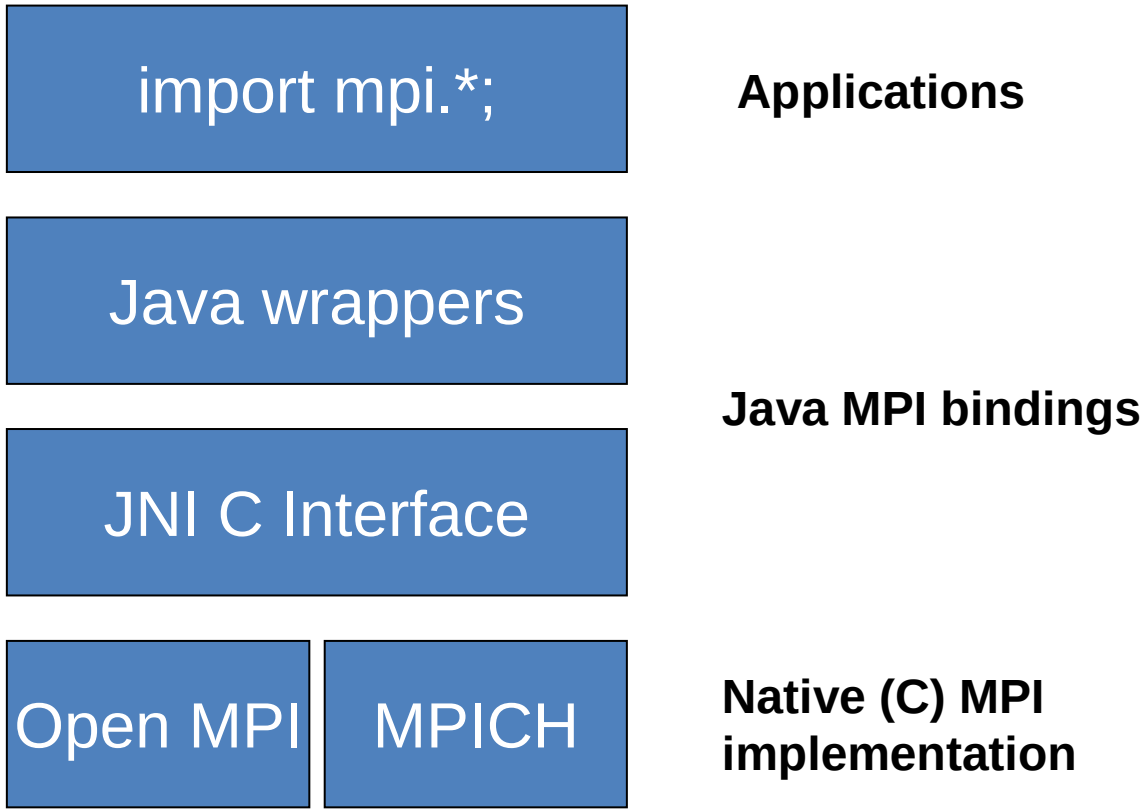
➢ Issues

lack of implementations for the programming languages typically used in the data-centric communities, such as Java

- **Java implementations** (MPJExpress)
  - full MPI-2 standard implementation
  - issues with supporting new high-speed interconnects (e.g. Cray), related to the JVM
  - scalability to a peta/exaflop?
  - support of native tools for parallel computing, i.e. debuggers, error detectors, etc.
- **native implementations** (mpiJava)
  - JNI is used for calling communication libraries that are available in native codes (i.e. highly optimized MPI comm.)
  - integration with a native MPI library is not easy
  - …but if you got it running, very enjoyable performance

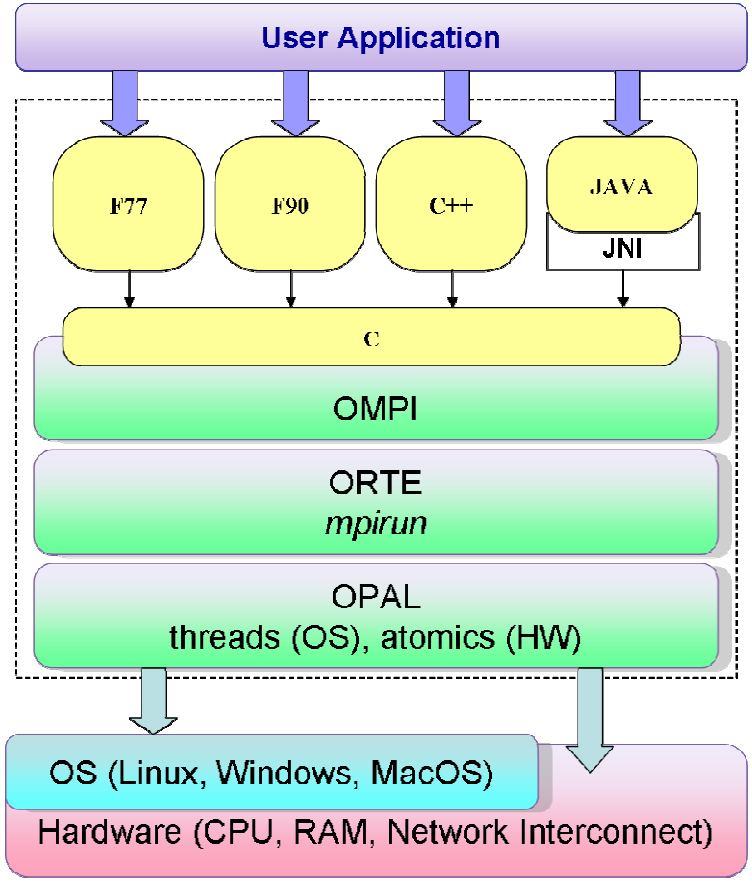
The Message-Passing Interface

- mpiJava
  - **Architecture**



The Message-Passing Interface

➢Java bindings for Open MPI (ompiJava)



configure –**enable-mpi-java**
make install
Configuration, installation

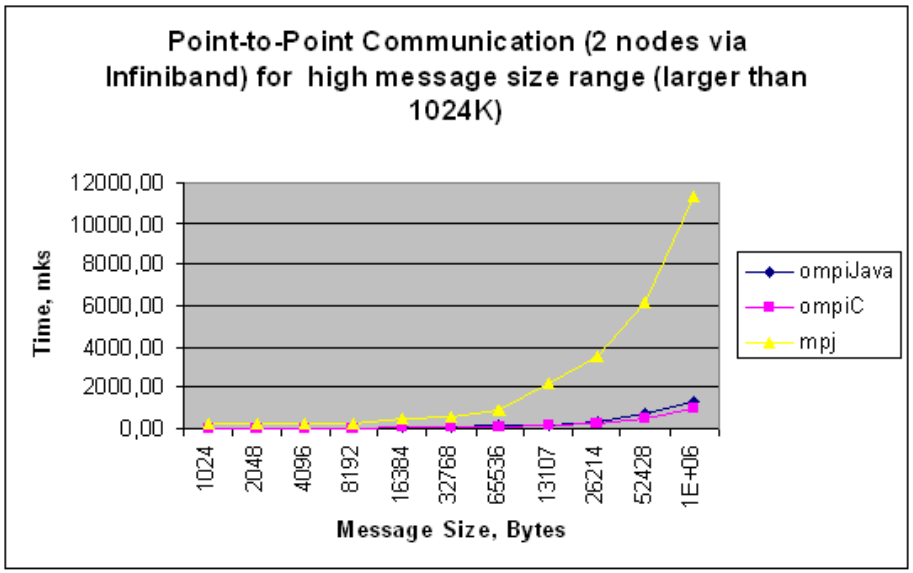
mpijavac *Application.java*
Compilation

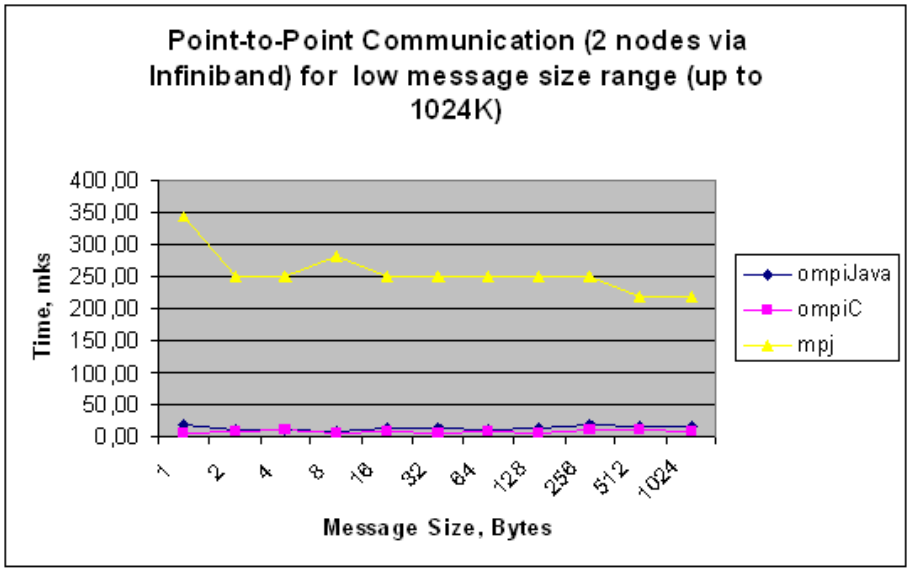
mpirun *java Application*
Running

Developments @ HLRS

➢ ompiJava performance

- P2P communication

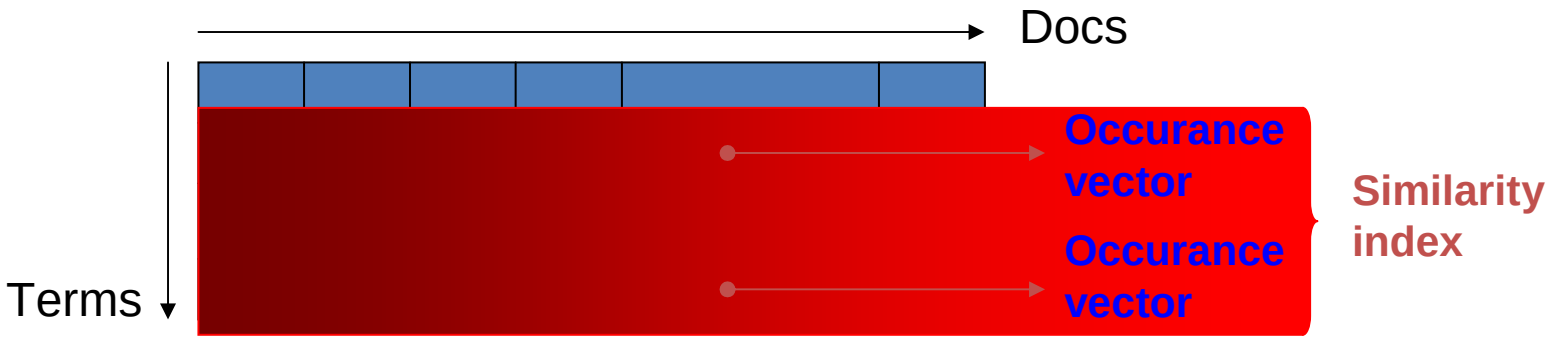




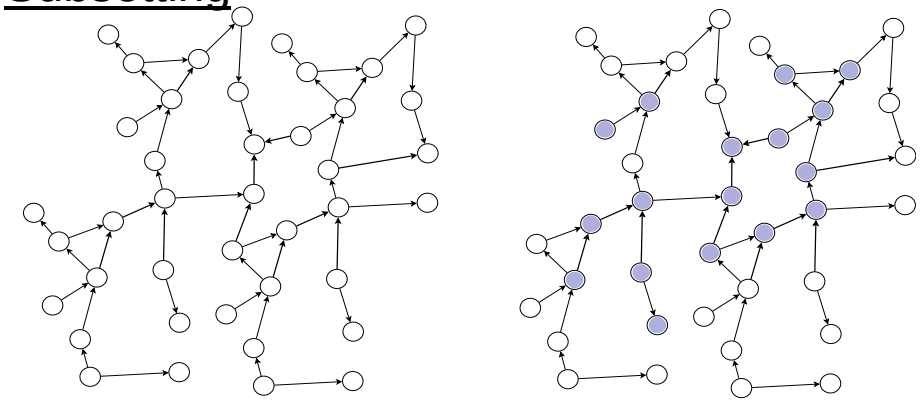
Application Scenarios

➢ Random Indexing for Large Texts

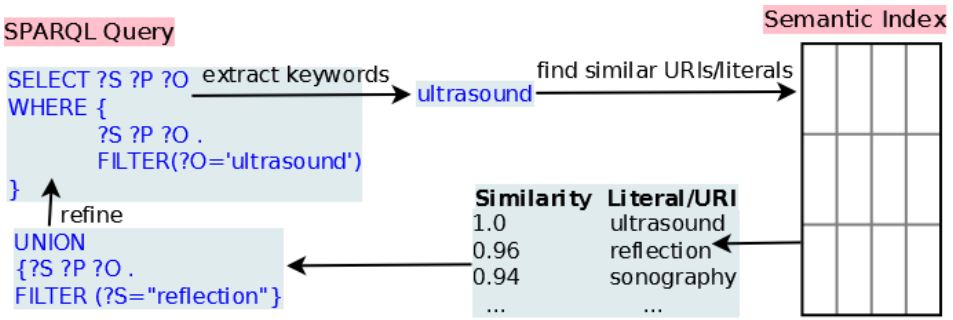
A Statistical Distribution technique for word/text similarity analysis



Subsetting



Query Expansion

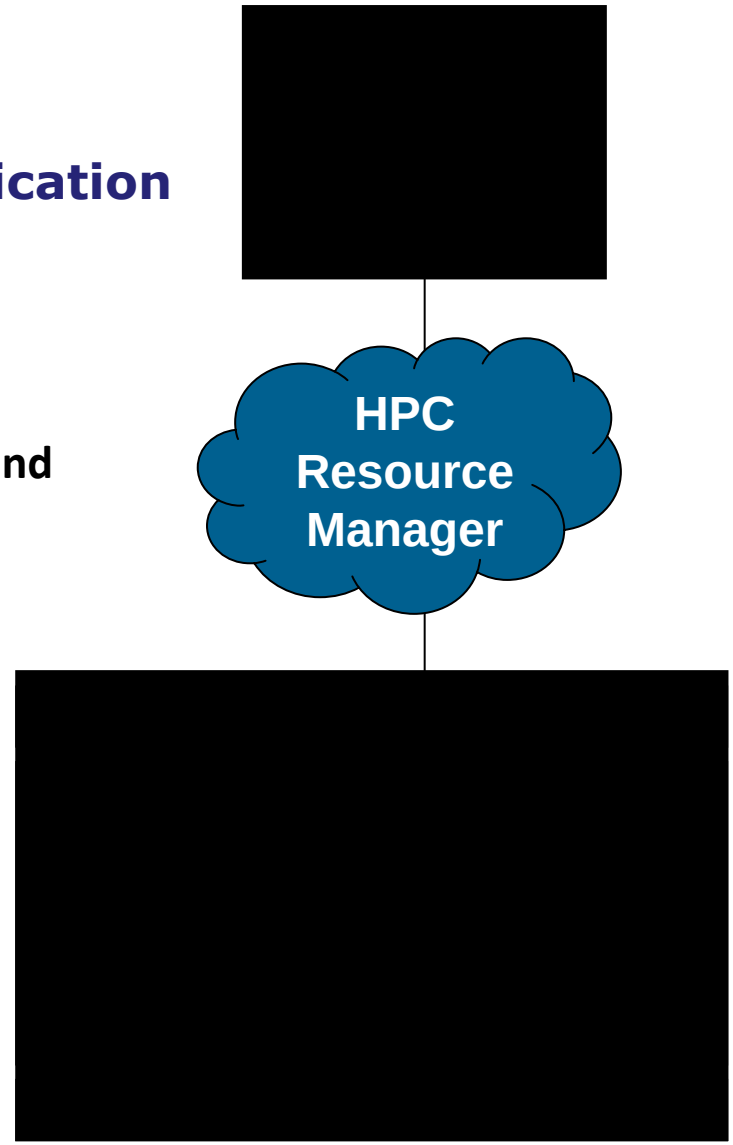


OUTLINE

- **About us**
- **Convergence of Supercomputing into Big Data**
- **Semantic Web as a new HPC domain?**
- **Mission Data-Centric Parallel Programming Models**
- **DreamCloud project approach**

5 DreamCloud Project Approach

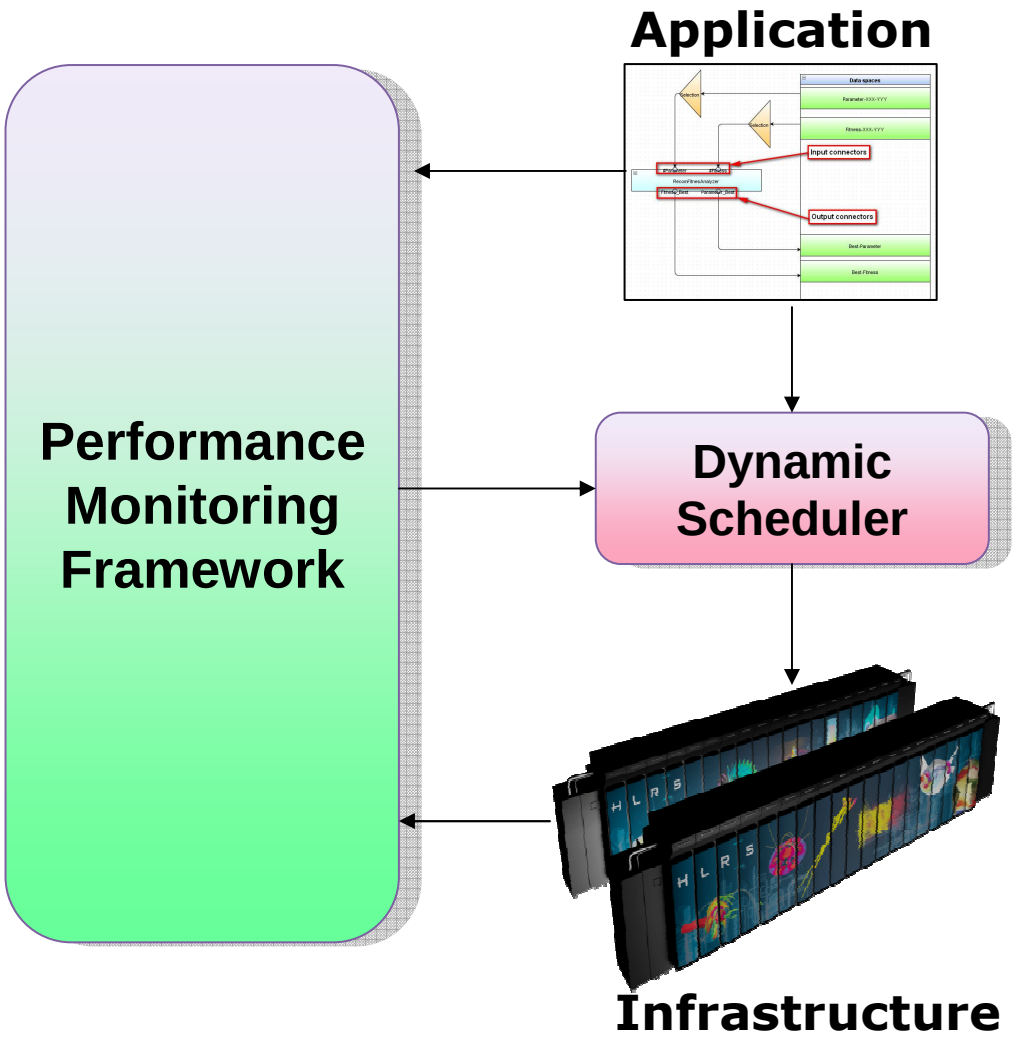
- Problem statement



- **I/O bound**
- **Communication-bound**

5 DreamCloud Project Approach

- DreamCloud solution
  - Scalable
  - Low overhead
  - Flexible
  - Unified



5 DreamCloud Project Approach

- Available application profiling tools
  - Valgrind – general analysis
  - Likwid – energy/power consumption
  - Vampir/Paraver – communication
  - Wireshark – networking infrastructure
  - …
- Infrastructure Monitoring Frameworks
  - Zabbix
  - Nagious
  - Excess
  - …
- **Problem: A very high user‘s involvement**
- **Key solution: Consolidation and Integration!**

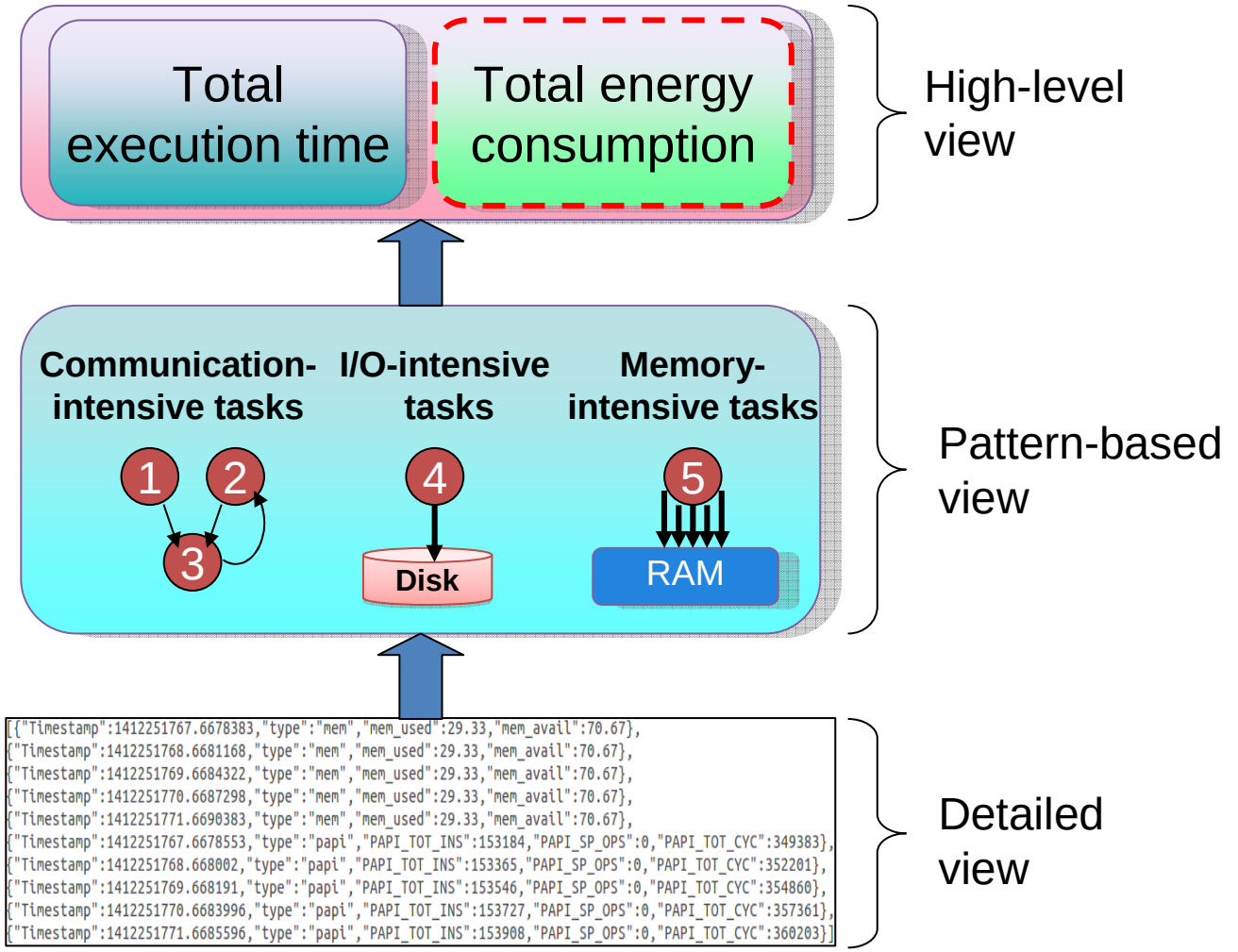
5 DreamCloud Project Approach

- Consolidated View

Legenda:

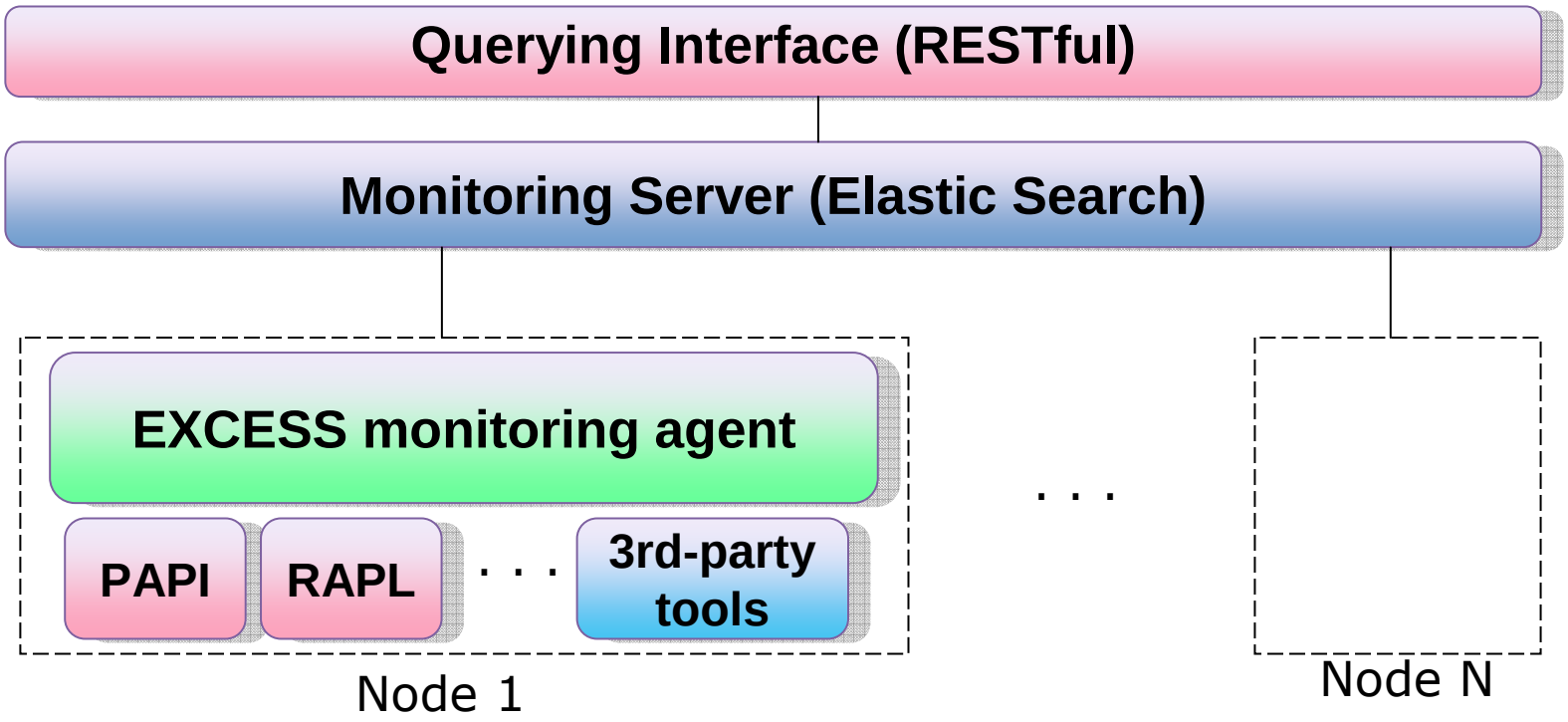
communication streams

to be targeted by D3.2



5 DreamCloud Project Approach

- Architecture



- Performance Profiles

  - Static (prior to the execution)
  - Dynamic (collected at runtime)
    - Performance Counters (PAPI etc.)
    - Energy Counters (RAPL etc.)
    - User-defined ones
      (e.g. progress tracker)

5 DreamCloud Project Approach

H L R S

- Basic Metrics

Metric	Value range	Description
PAPI_L1_DCM	[integer]	Level 1 data cache misses
PAPI_L1_ICM	[integer]	Level 1 instruction cache misses
PAPI_L2_DCM	[integer]	Level 2 data cache misses
PAPI_L2_ICM	[integer]	Level 2 instruction cache misses
PAPI_L1_TCM	[integer]	Level 1 cache misses
PAPI_L2_TCM	[integer]	Level 2 cache misses
PAPI_L3_TCM	[integer]	Level 3 cache misses
PAPI_TLB_DM	[integer]	Data translation lookaside buffer misses
PAPI_TLB_IM	[integer]	Instruction translation lookaside buffer misses
PAPI_L1_LDM	[integer]	Level 1 load misses
PAPI_L1_STM	[integer]	Level 1 store misses
PAPI_L2_STM	[integer]	Level 2 store misses
PAPI_STL_ICY	[integer]	Cycles with no instruction issue
PAPI_BR_UCN	[integer]	Unconditional branch instructions
PAPI_BR_CN	[integer]	Conditional branch instructions
PAPI_BR_TKN	[integer]	Conditional branch instructions taken
PAPI_BR_NTK	[integer]	Conditional branch instructions not taken
PAPI_BR_MSP	[integer]	Conditional branch instructions mispredicted
PAPI_BR_PRC	[integer]	Conditional branch instructions correctly predicted
PAPI_TOT_INS	[integer]	Instructions completed
PAPI_FP_INS	[integer]	Floating point instructions
PAPI_LD_INS	[integer]	Load instructions

Metric	Value range	Description
CACHE_ACCESSES CACHE_MISSES	[integer]	Number of (e.g. L1, L2, data, instruction) cache accesses and misses. If ratio of misses/access is too high a different platform, featuring bigger caches, or a more fine grained work distribution, such that each process' workload fits into the local cache, could be beneficial.
INSTRUCTIONS_RETIRED	[integer]	The amount of instructions retired (completed) per clock cycle indicates bottlenecks (e.g. ops waiting for data from memory). Possible solutions include that alleviate this problem include using Hyperthreading on a node or utilization of specialized program version (e.g. optimized for SSE or AVX instructions).
FLOATINGPOINT_OPS ARITHMETIC_OPS	[integer]	A high amount of floating point/arithmetic operations per clock cycle coupled with a low number of retired operations could be solved by using a different architecture/platform or program version (e.g. changing platform to avoid shared floating point units on AMD Bulldozer architecture).
ENERGY_CONSUMPTION CORE_TEMPERATURE THERMAL_SPECS	[integers]	The RAPL interface allows reading of several energy consumption and thermal readings as well as the hardware specifications. This information can be used to select the best
PAPI_STL_ICY	[integer]	Cycles with no instruction issue
PAPI_BR_UCN	[integer]	Unconditional branch instructions
PAPI_BR_CN	[integer]	Conditional branch instructions
PAPI_BR_TKN	[integer]	Conditional branch instructions taken
PAPI_BR_NTK	[integer]	Conditional branch instructions not taken
PAPI_BR_MSP	[integer]	Conditional branch instructions mispredicted
PAPI_BR_PRC	[integer]	Conditional branch instructions correctly predicted
PAPI_TOT_INS	[integer]	Instructions completed
PAPI_FP_INS	[integer]	Floating point instructions
PAPI_LD_INS	[integer]	Load instructions

5 DreamCloud Project Approach

- Power consumption

Channel id	Board id	Node	Component	Power characteristic
0	0	node01	CPU1+RAM	A
1	0	node01	CPU1+RAM	V
2	0	node01	CPU2+RAM	A
3	0	node01	CPU2+RAM	V
4	0	node01	PCIe	A
5	0	node01	ATX (+12 volt)	A
6	0	node01	GPU1	A
7	0	node01	GPU1	V
8	1	node02	CPU1+RAM	A
9	1	node02	CPU1+RAM	V
10	1	node02	CPU2+RAM	A
11	1	node02	CPU2+RAM	V
12	1	node02	PCIe	A
13	1	node02	ATX (+12 volt)	A
24	3	node01	Entire device	Watt
25	3	node02	Entire device	Watt
26	3	frontend	Entire device	Watt
27	3	NAS	Entire device	Watt
28	3	Infniband switch	Entire device	Watt
29	3	Ethernet switch	Entire device	Watt

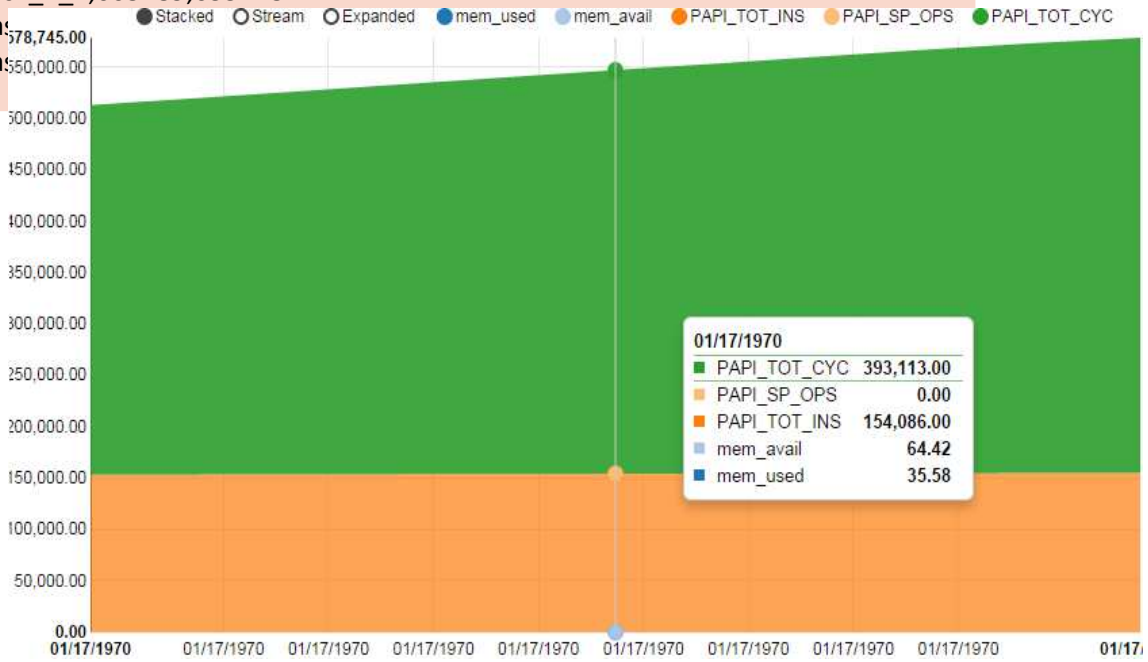
5 DreamCloud Project Approach

- Performance Profiles Representation in DreamCloud

Performance Profiles (Example in CSV)

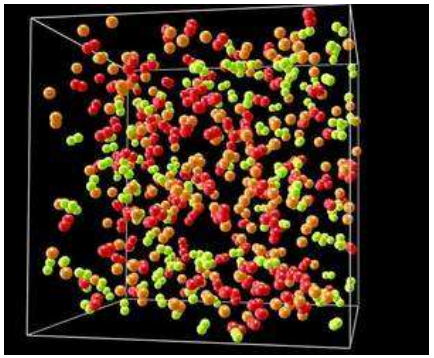
```
timestamp;node;task;PAPI_TOT_INS:CPU0;PAPI_TOT_INS:CPU1
1418120193.4880380;node_1[perf];task_1;516680185;663789
1418120194.4996216;node_2[perf];task_1;663789;663714
1418120196.5121722;node_3[balanced];task_1;3805840;663405
1418120193.4880380;node_3[energy];task_2_1;516680185;623687
1418120194.4996216;node_9[perf];task_2_1;663789;655418
1418120196.5121722;node_3[perf];ta
1418120194.4996216;node_9[perf];ta
```

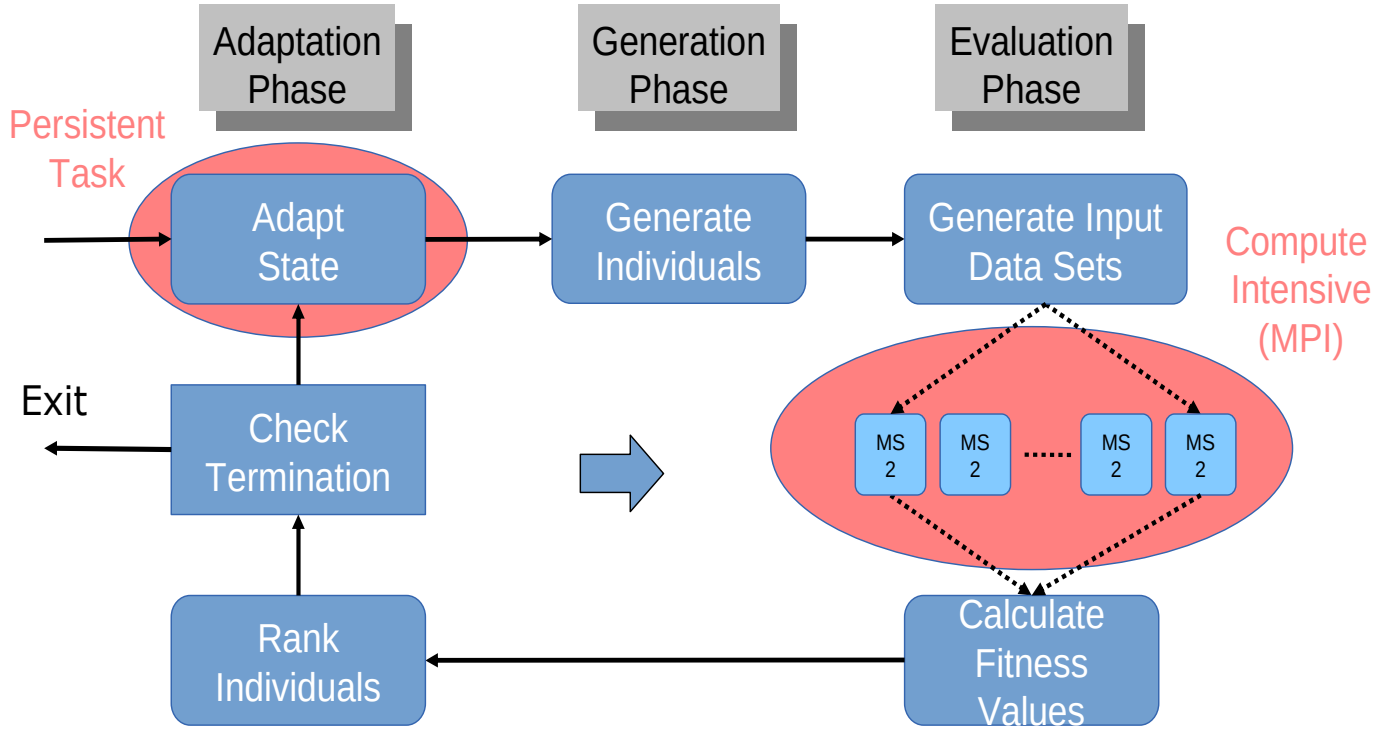
- Visualization



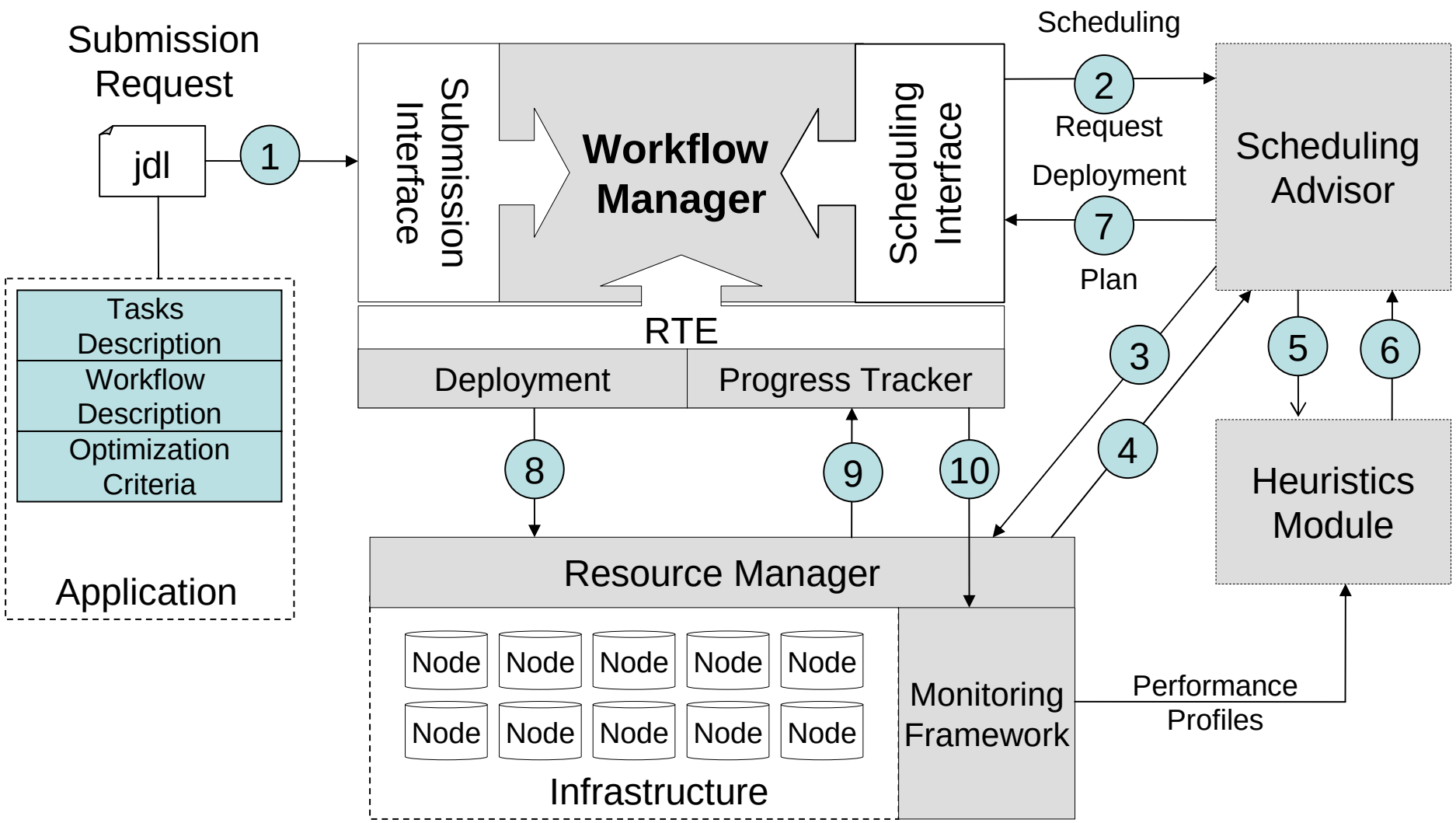
5 DreamCloud Project Approach

- Molecular Dynamics (MD) Simulation Code – MS2
  - Simulation of movement of atoms and molecules
  - Massively parallel, on a per-particle basis





5 DreamCloud Project Approach



6 Conclusion

Main Results

- HPC is going to face new challenges related to data-centric application expansion.
- Parallel programming models (mainly MapReduce and MPI) are the key enablers of HPC to data-centric applications
- Reaching near-peak performance is going to be the major challenge

Future Work

- Promote existing technologies, such as MPI, to solving new challenges, such as Big Data.
- Making existing framework more data-centric.