NESUS COST IC1305
Network for Sustainable Ultrascale Computing

nesus.eu

Network for Sustainable Ultrascale Computing

COST Action IC1305

Russian Supercomputing Days
Moscow, 26-27 September 2016

Prof. Jesus Carretero
Nesus Action Chair
University Carlos III of Madrid
Spain

arcos.inf.uc3m.es

Contents

- University Carlos III of Madrid-ARCOS
- NESUS: Network for sustainable ultrascale computing
- Ultrascale storage I/O stack

University Carlos III of Madrid

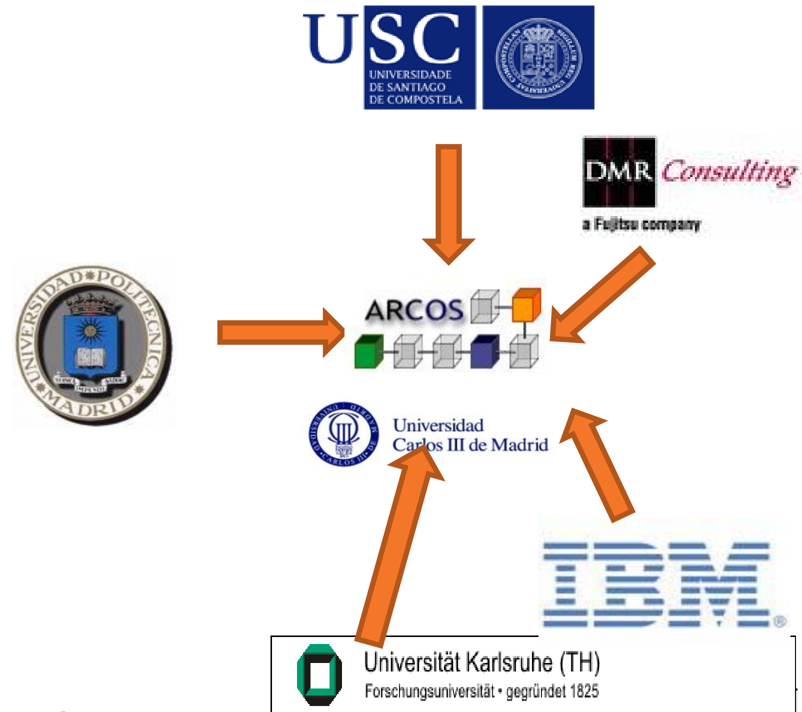
- Created in 1989.
  - 25,000 students
- Centers:
  - Social Sciences and Law School
  - Humanities and Journalism School
  - Engineering School.
    - Computer Science & Engineering Department
      - Research group:
        Computer Architecture and Systems (ARCOS)

Leganés

Madrid, Spain

ARCOS Research Group

- Created in 1999.
  - Leader: Jesus Carretero
- Staff:
  - 2 Professors
  - 4 associate professors
  - 3 assistant professors
  - 5 researchers
  - 12 PhD Students.



- Goals:
  - Applied research on large-scale parallel and distributed systems (parallelization, runtimes and I/O).
- Contacts:
  - Argonne Labs, Northwestern, CINVESTAV, DKRZ, INRIA, CNRI, CIBERSAM, IBM Research, …

Contents

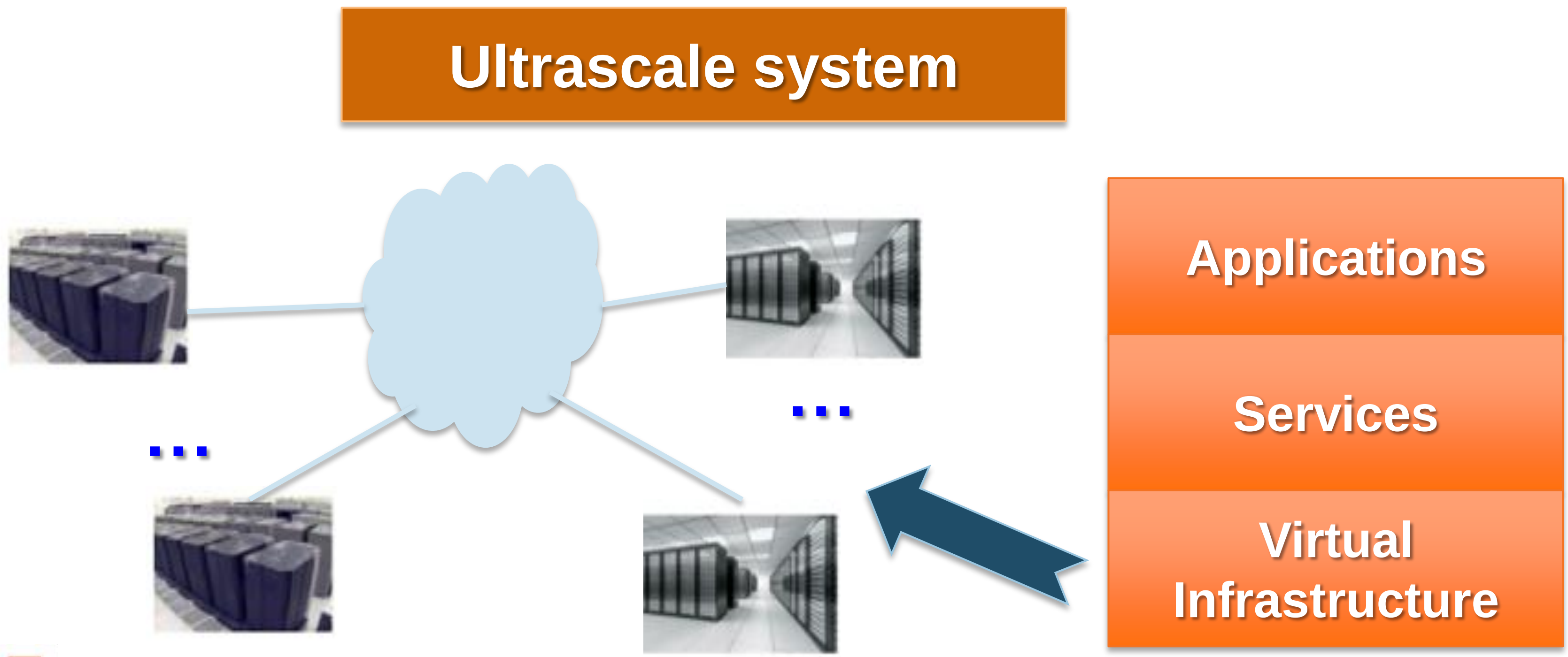
- NESUS: Network for sustainable ultrascale computing
- Ultrascale storage I/O stack

Current scenario

- More complex computing scenarios
  - HPC, HTC, MTC, DIC, ..
  - With different requirements
- There are major research efforts towards:
  - Exascale (PRACE, EESI, HP-SEE, IESP)
  - Large scale virtual systems (XSEDE, FutureGrid, Grid5000).
  - Big data solutions (BIG, EIOW, BDEC)
- Efforts are mostly separated
  - But convergence is needed, and required by users.

Ultrascale systems

- Ultrascale computing systems (UCS)
  - Big-scale complex system integrating parallel and distributed computing systems, that cooperate to provide solutions to the users at unprecedented scale.



Promote sustainability

- As the scale and complexity increase in UCS, **sustainability is becoming a major challenge**

- Sustainability not only means energy, but all factors that will allow the system to be adopted and maintained.

- Sustainability in UCS should be the result of leveraging several cross-layer aspects to face complexity:
  - Programmability, Data management, Resilience, Energy efficiency, Scalability, …

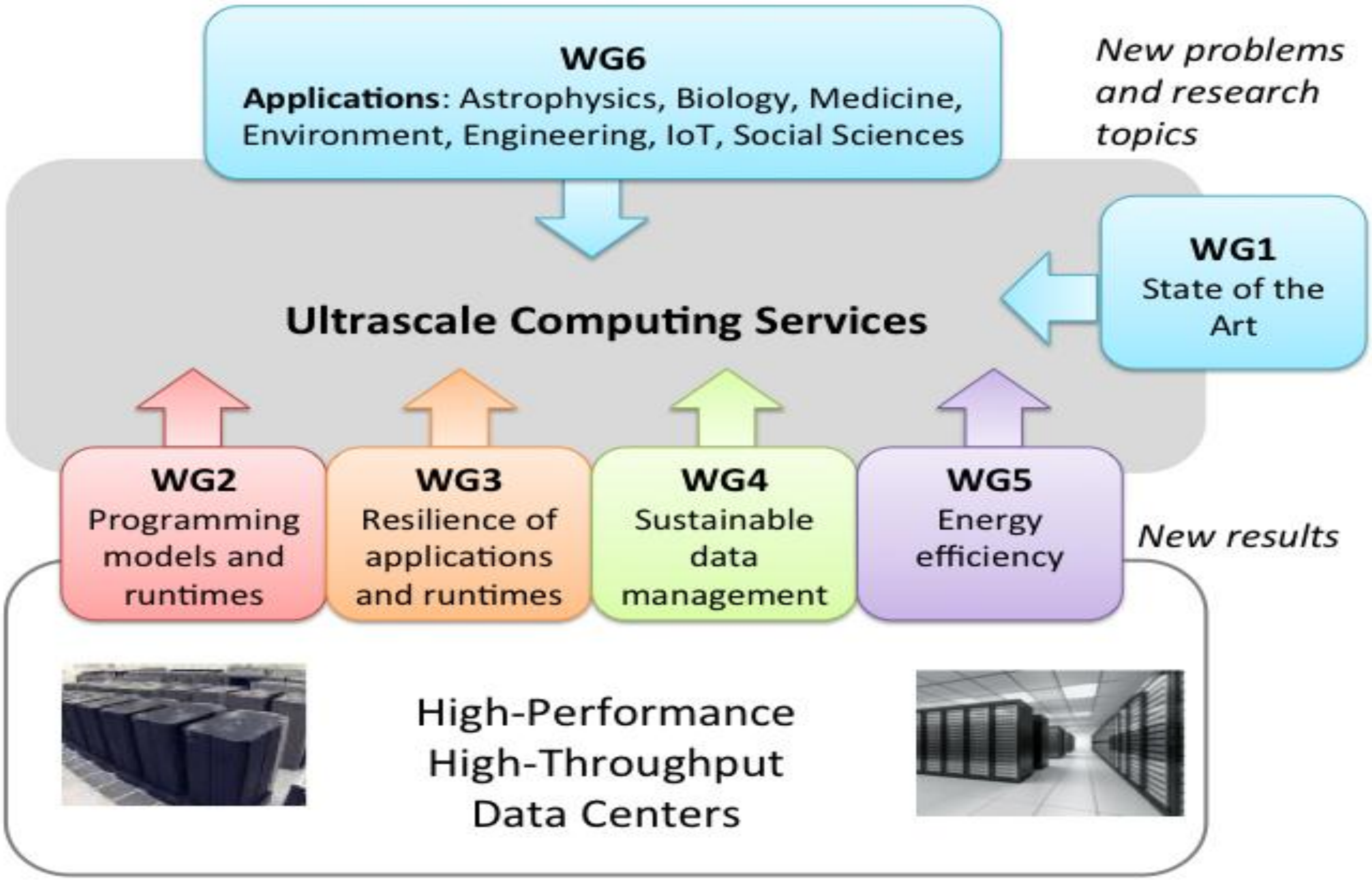
Scientific goals

- Exploring new solutions for the **system software stack** (programming paradigms, runtimes, middlewares, resilience, data management, and energy models) and their application to enhance sustainability in UCS.
  - Understanding trade-offs and synergies to leverage all factors.
  - Considering new hardware and architectural solutions.

- Exploring **redesign and reprogramming** efforts for applications to efficiently exploit ultrascale platforms, while providing sustainability.

- Holistic approach to **manage the whole ecosystem**,
  - Important to understand how all the factors affect UCS sustainability -> sustainability metrics

Workplan

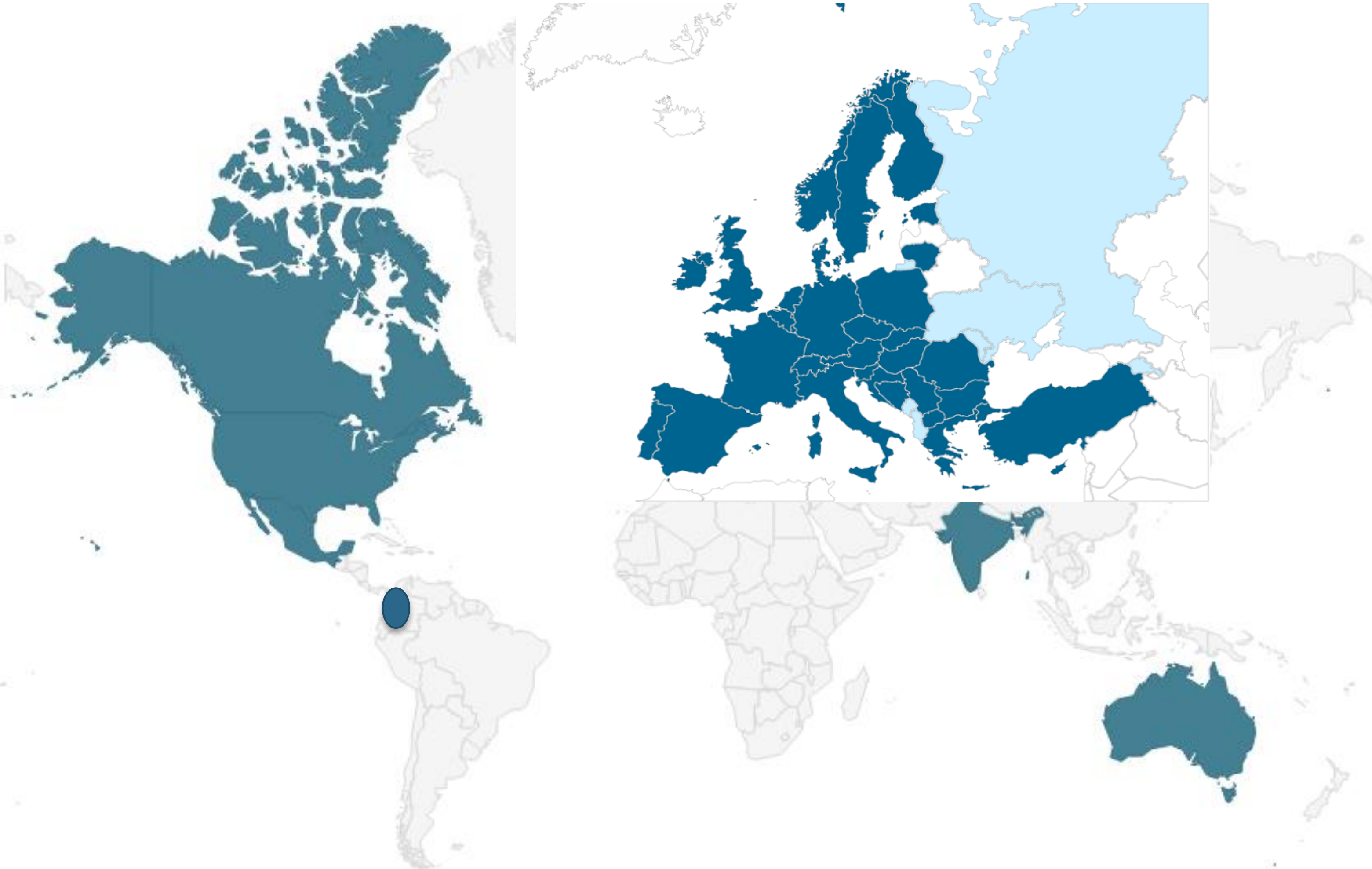

WG6
Applications: Astrophysics, Biology, Medicine, Environment, Engineering, IoT, Social Sciences
New problems and research topics
Ultrascale Computing Services
WG1
State of the Art
WG2
Programming models and runtimes
WG3
Resilience of applications and runtimes
WG4
Sustainable data management
WG5
Energy efficiency
New results
High-Performance
High-Throughput
Data Centers

NESUS Activities

- Working Group meetings
- Research stays -> 15 grants per first year
- Action workshop (2015 in Krakow, 2016 in Sofia).
- Winter school & PhD Symposium (2016 in Timisoara, 2017 Calabria)

- Strong emphasis in cooperation: open to external actors
  - Join publications, tools, applications, …
  - With industry to solve real-world cases
  - With other institutions/projects to advance in scientific goals

Consortium



45 countries

76 institutions

240 members

35% young researchers

Open to new members

Open to cooperation

Contents

- Ultrascale storage I/O stack

Applications I/O requirements

- Applications generate 10s of Tbytes of data per execution.

Project	On-line Data (TBytes)	Off-line Data (TBytes)
Laser-Plasma Interactions	60	60
Type Ia Supernovae	75	300
Lattice Quantum Chromodynamics	300	70
Engineering Design of Fluid Systems	3	200
Multi-material Mixing	215	100
Earthquake Wave Propagation	1000	1000
Fusion Reactor Design	50	100

- Keeping hundreds of Tbytes of data online is increasingly common.

Current problems of I/O stack

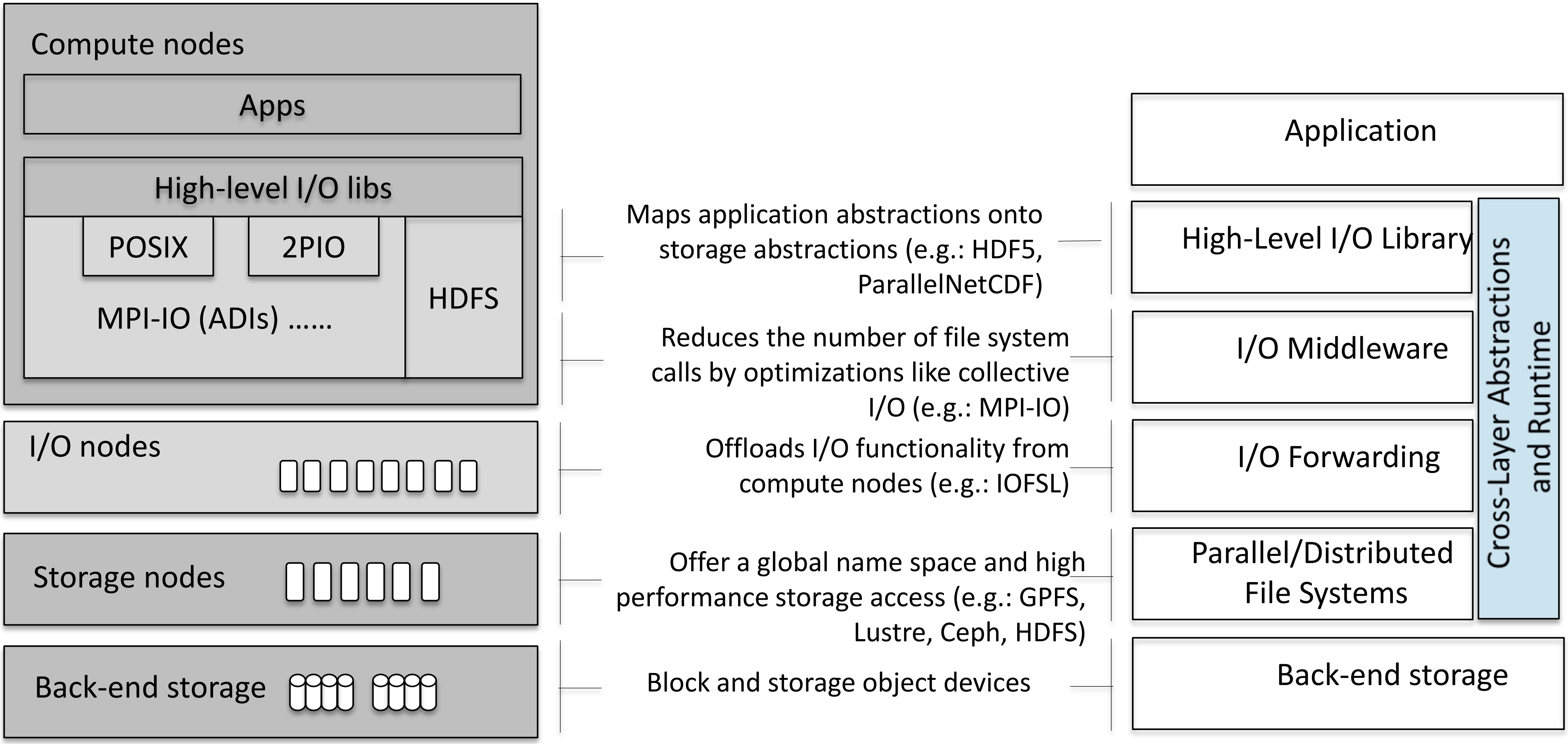
- As applications grow
  - Large scale data sets and conflicting data distribution models
- As the depth of the storage hierarchy increases
  - Programmability, performance, and data management are big concerns.
- I/O system optimizations applied independently at each system layer
  - Can cause mismatches between different layers
- Lack of mechanisms for adapting to unexpected conditions
  - Cross-layer adaptive control mechanisms not available for UCS I/O stack.
  - I/O interfaces are rigid and cannot be extended with new services over the data.
- Lack of capability of exposing and exploiting data locality
  - Dynamic deployment of I/O system not easy
- FS scalability is limited
  - Mostly due to metadata (~ 65% ops)

First step: coordinating functionality at I/O stack levels

- Vertical coordination ->
  - Mapping application models on storage models
  - Coordinate multiple level buffering/caching for latency hiding
  - Controlling vertical data flow: CN <-> ION<-> FS <-> SN

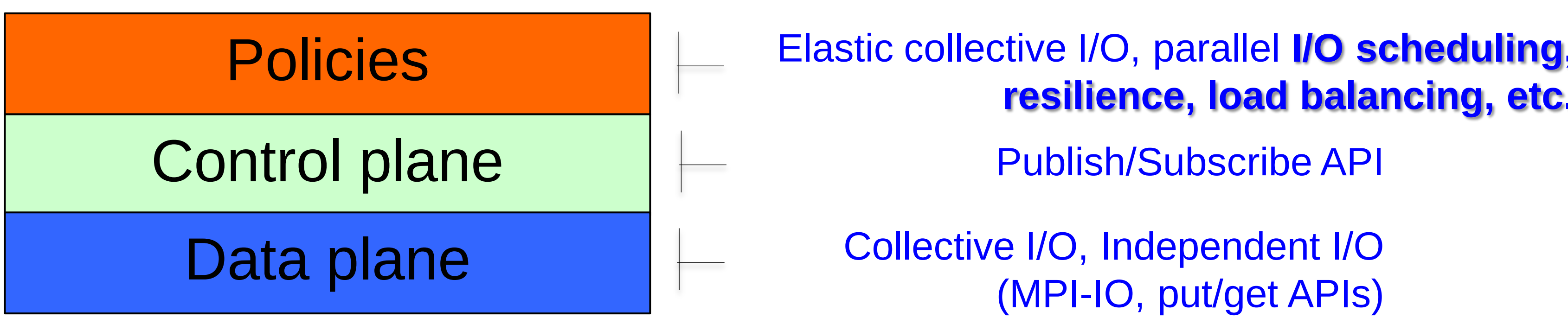
- Horizontal coordination ->
  - Transparent access to distributed (unstructured) data
  - Collective I/O on compute nodes
  - (In-memory) data aggregation on the I/O nodes
  - Transparent replication of data

Proposal: CLARISSE



CLARISSE Architecture

- Cross-layer abstractions at run-time
  - Facilitate the flow of control and data across the I/O stack
- Decouple the data and control planes
  - Data plane
  - Control plane
  - Policies plane

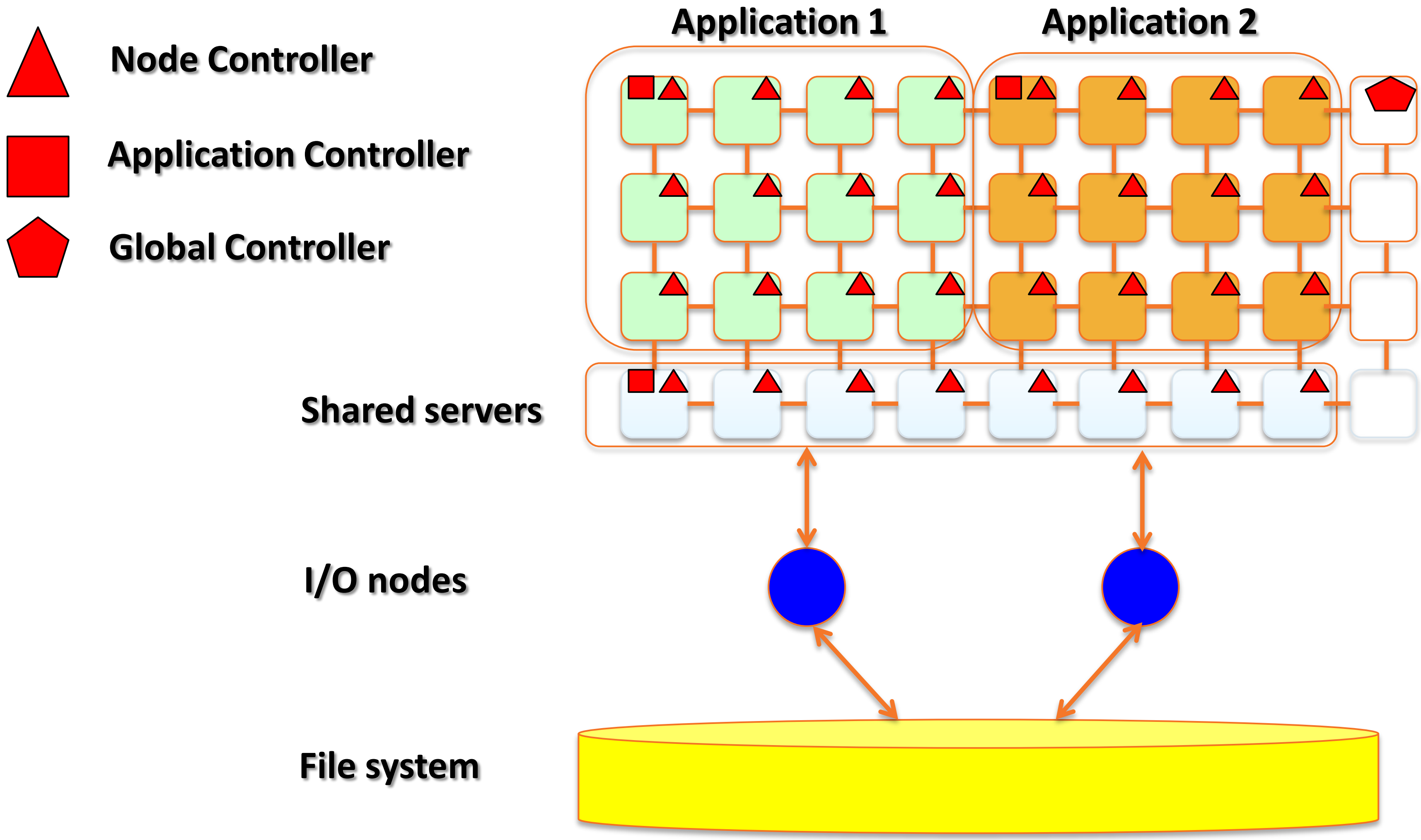


CLARISSE control plane

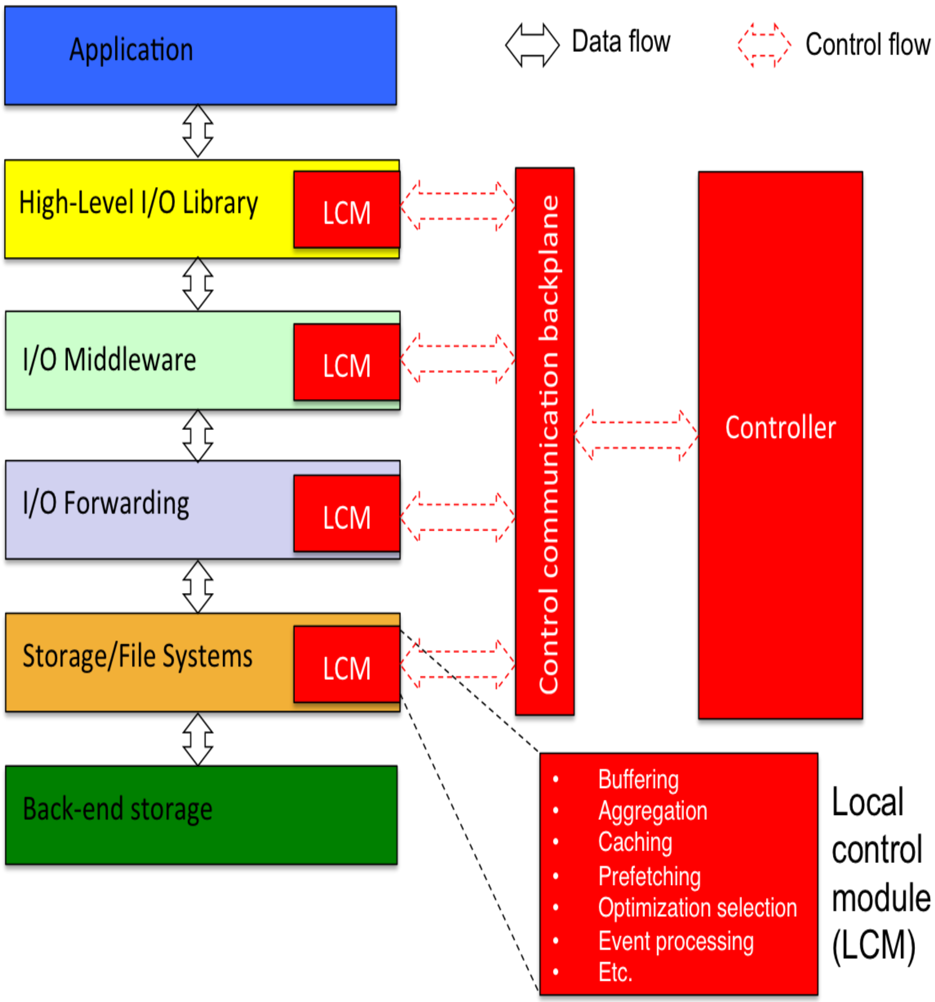
- Control backplane
  - Based on a publish/subscribe substrate (e.g. Beacon)
  - Processes can subscribe to events having certain properties
    - Associate call-back
    - Wait for an event
    - Check for the arrival of an event

- Allows building any distributed/replicated control architecture
  - All nodes participate in control

Example: hierarchical control infrastructure



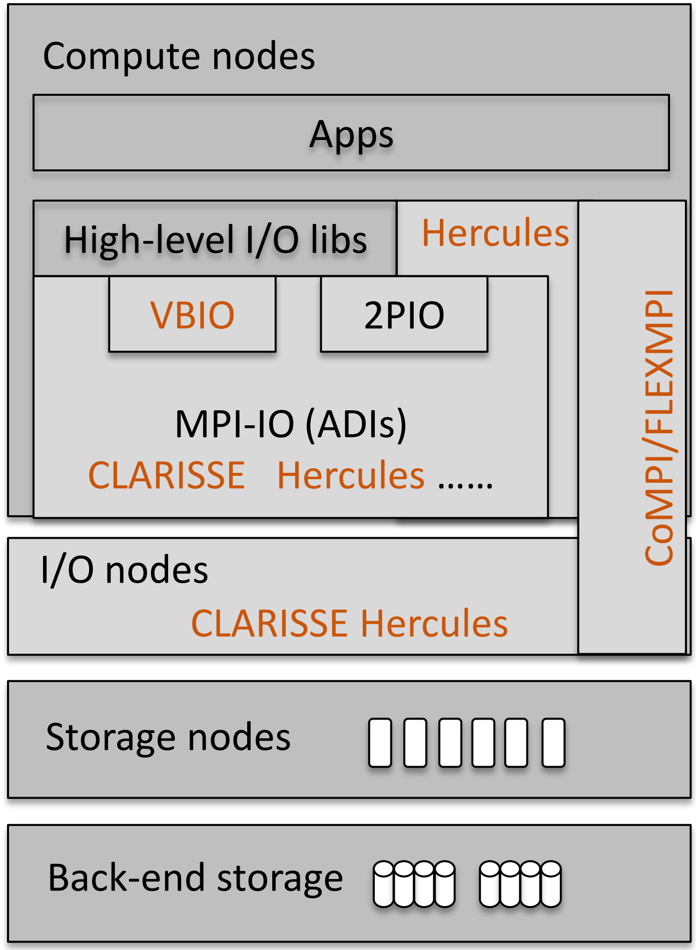
Data plane



- Design novel abstractions and mechanisms for supporting data flow optimizations
  - Data aggregation (e.g., collective I/O)
  - buffering / caching, data staging, in-memory
  - load balance
  - data locality (e.g. in-situ and in-transit data processing)
- Parallel data-flows based on the these abstractions

Data management components

- View-based I/O (VBIO)
  - File views I/O optimization for high performance collective file access
- Hercules
  - Dynamic deployment of in-memory object-stores per node
  - Guided by the scheduler
  - Put/Get API (Key-value)
- FlexMPI
  - Elastic deployment of processes and I/O servers

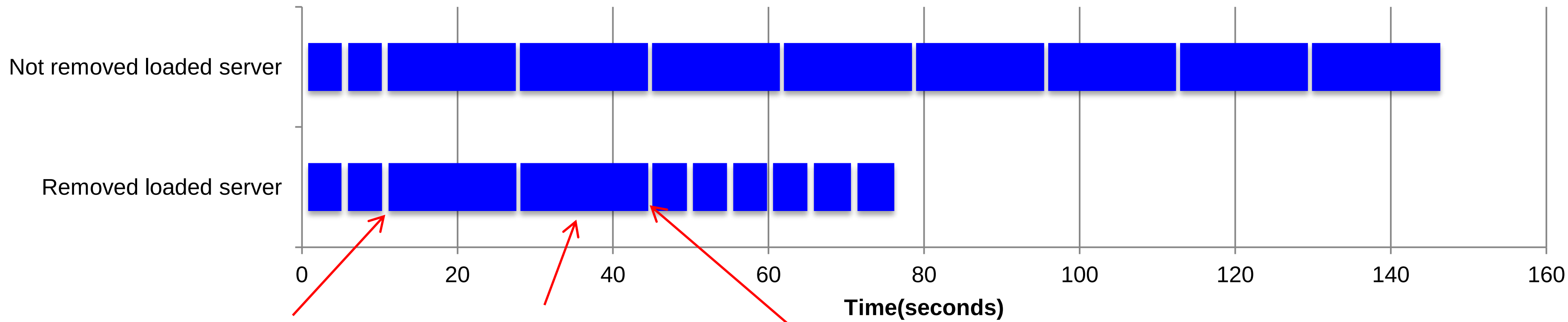


CLARISSE policies plane

- Decitions taken based on control and data info.
  - Data distribution and load balancing,
    - Data-location aware scheduling
  - Resilience,
    - Automatic replication
  - Elastic collective I/O,
    - Enhance large collective I/O operations
  - Parallel I/O scheduling,
    - Enhances scheduling of multiple parallel I/O operations
  - ….

Elastic collective I/O scheduling evaluation

Write time (10 operations, 3840 processes, 256/255 servers)

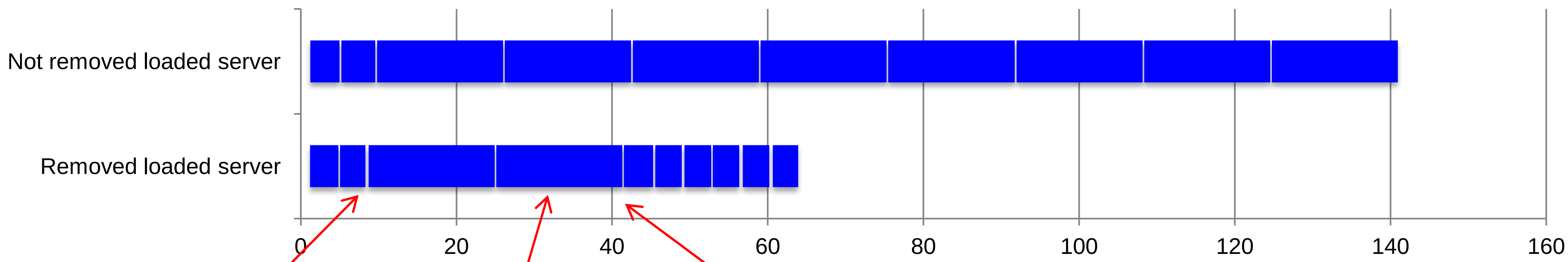


Loaded server

Detect loaded server
Reconstruct server map

New epoch with fewer servers

Write time (10 operations, 15360 processes, 1024/1023 servers)

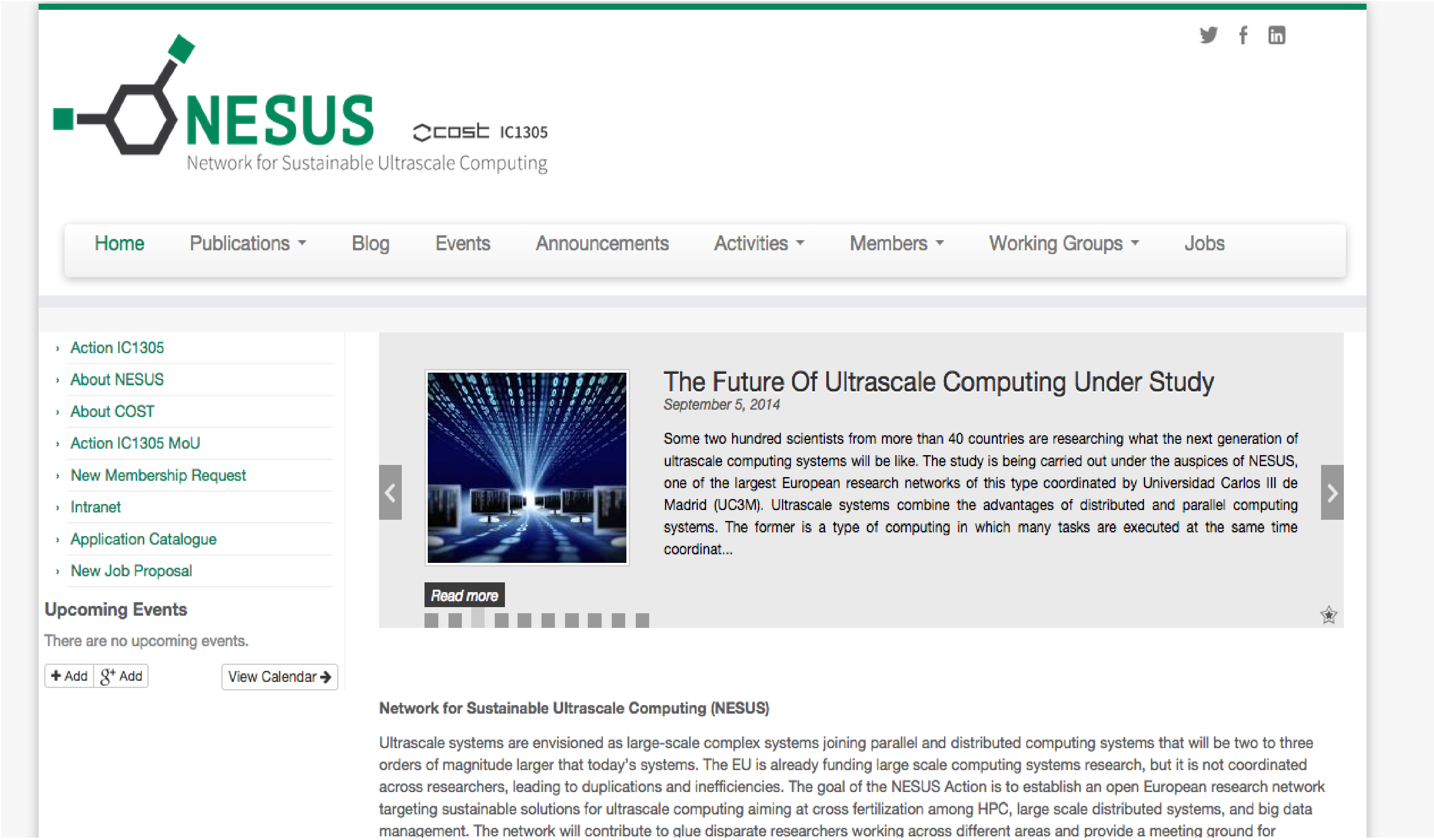


Loaded server

Detect loaded server
Reconstruct server map

New epoch with fewer servers

NESUS Web portal (nesus.eu)



NESUS cost IC1305 nesus.eu

Network for Sustainable Ultrascale Computing

ACM/IEEE CCGrid 2017

Madrid, Spain, May 14-17, 2017

See you in Madrid!

Thank you!

arcos.inf.uc3m.es