Clustrx: A new generation operating system designed for HPC

HPC 2010, June 21-25, Cetraro

Dmitry Tkachev, R&D Director

T-PLATFORMS

www.t-platforms.ru

The Petascale Challenge

MSU Installation: 0.5PFlop ~ 5000 nodes

For installation of 10-100 PetaFlop:

~10000 or 100000 or 1 million nodes

…as a single resource

What challenges must the Operating System address?

T-PLATFORMS

www.t-platforms.ru

The 10 plagues of OS

1. Reliability: Constantly changing configuration of the system
2. Resource management: What nodes/processors can be used ?
3. Management of the nodes: Switch on/off
4. Monitoring: What's the temperature? What is used/not-used?
5. Electricity: can parts be switched off?
6. File System: Where is the data?
7. Provisioning: What version of Linux/Windows/etc is/should install
8. Remote load: how to load the nodes in short time?
9. Keeping track of the

The 10th plague

10. Does it scale?

Solution definition

The node Operating System (Linux, SUSE, Red Server, Cray OS, etc...)

Is NOT your solution
(local resource allocation and utilization)

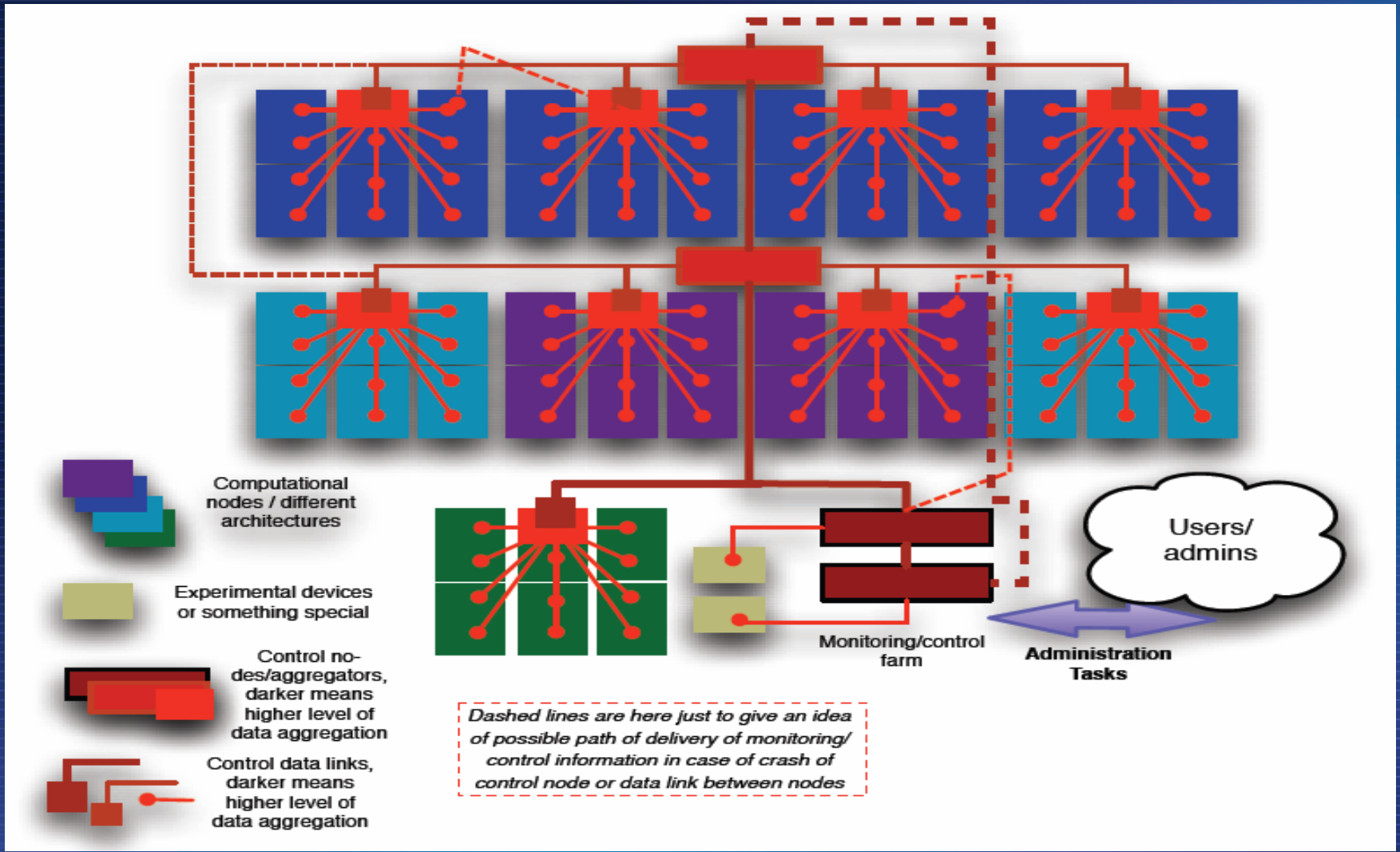
System Management software is the solution:

The (Cluster) Operating System

Clustrx – The (Cluster) Operating System

- **Light weight real time monitoring of ALL resources**
  - **Computational nodes, switches, electricity, cooling, etc**
- **Clustrx is responsible for full life cycle of the system:**
  - **Emergency switching on/off on particular failures**
  - **Switching off to conserve electricity**
- **Resource allocation for computation**
  - **Match of program needs to available**

Clustrx's architecture



Clustrx Subsystems

1. Clustrx Watch **- monitoring and control**
1. dConf - **Cluster-wide, decentralized distributed storage for configuration data**
1. Resource manager - **POSIX-compliant, modular, scalable, GRID-ready**
1. Network boot & provisioning - **infrastructure to support any number of computing nodes**

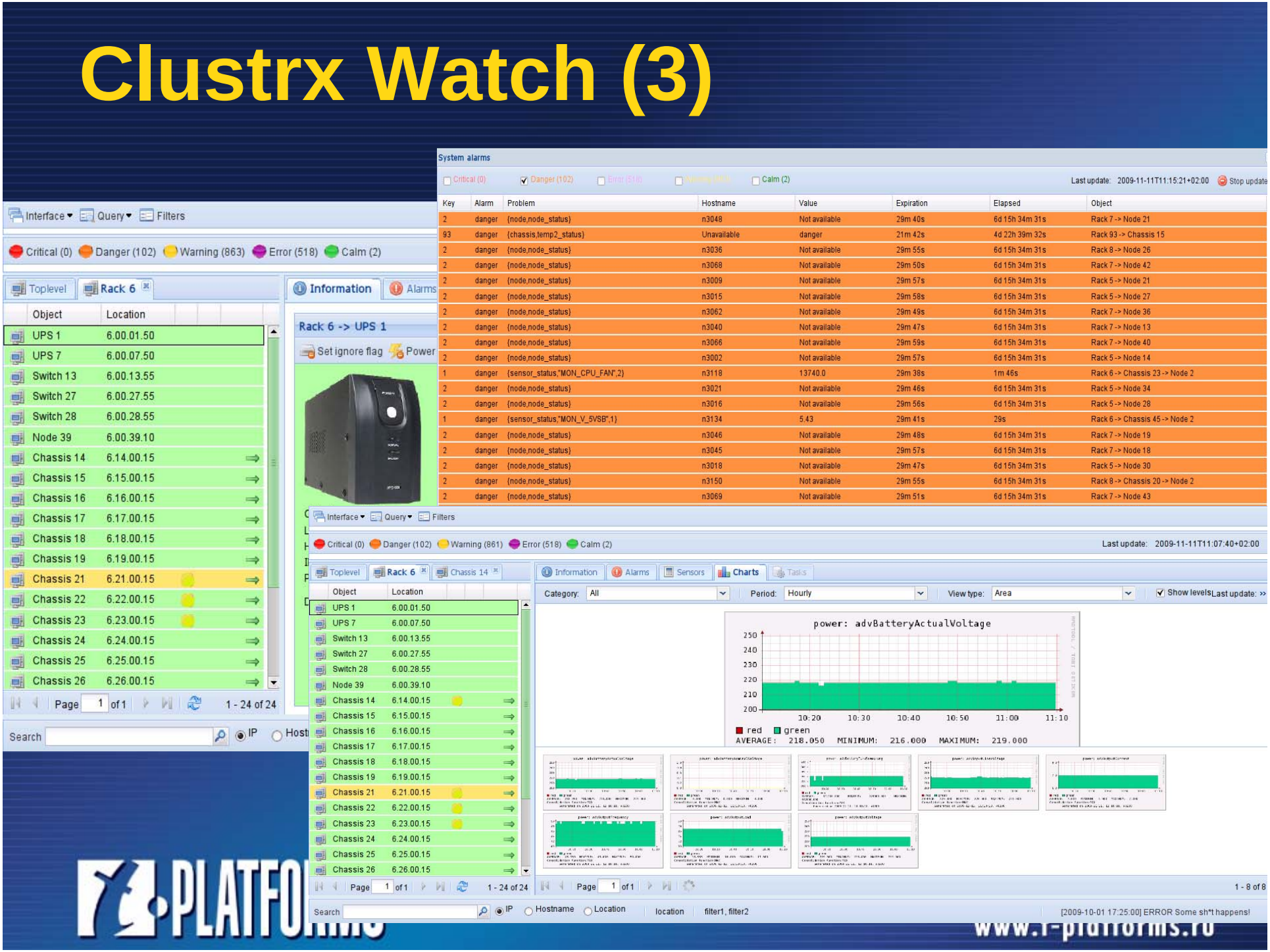
1. Clustrx Watch

- Near real-time monitoring of 10000's of nodes (*no practical limit*)
- Full support of heterogeneous architectures
- Monitoring agents for all major OSes with low CPU load
- IPMI monitoring and management
- SNMP monitoring
- Cluster state visualization
- Automated actions (including emergency shutdown)
- 3rd parties' SW monitoring through open API
- Adaptive schemes of task management and cluster controls

Clustrx Watch (2)

Cluster state visualization

Clustrx Watch (3)



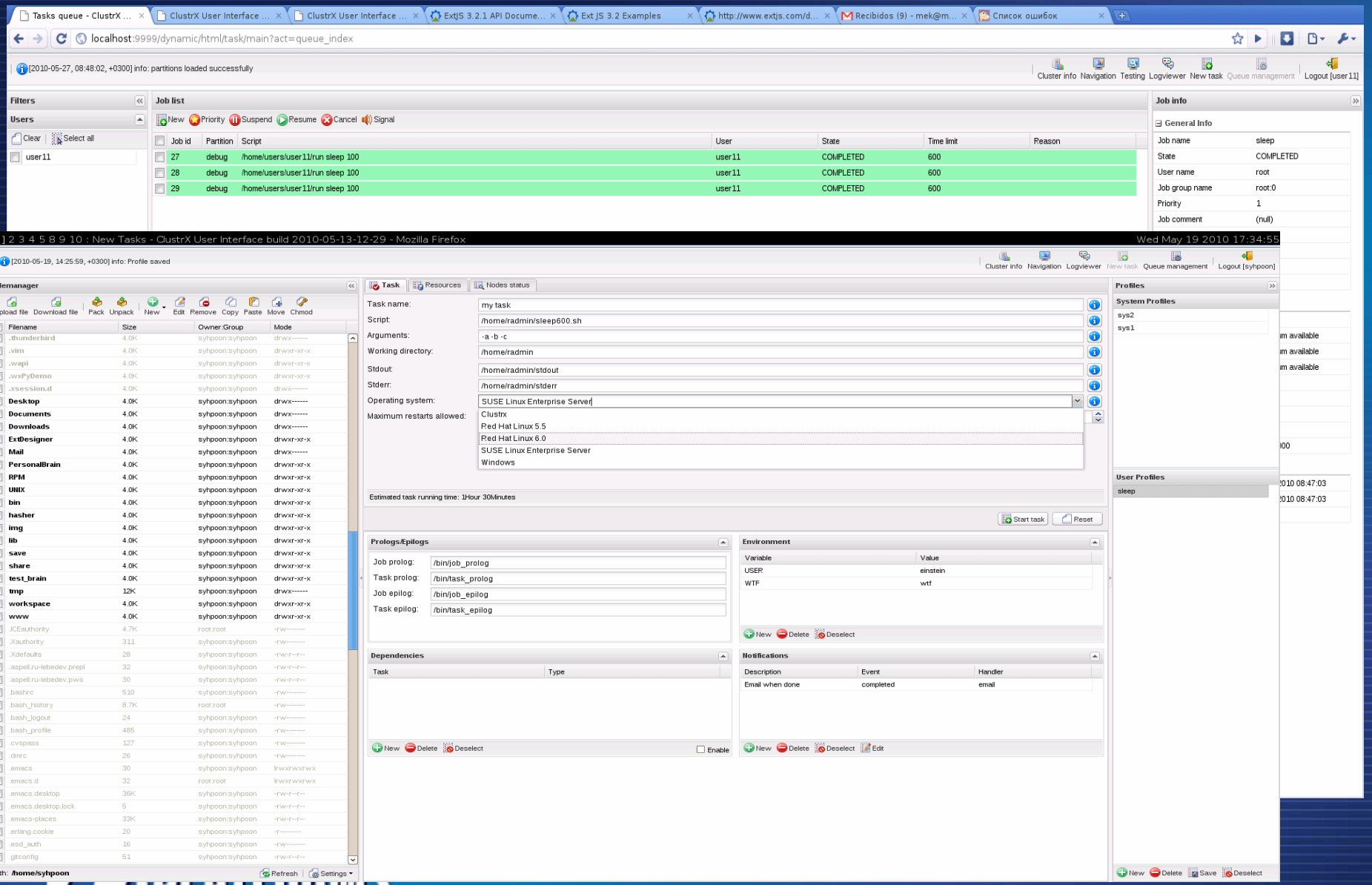
2. Clustrx dConf

- Distributed scalable database to keep configuration data
- Tight integrated with other subsystems
- Unified access to configuration data for systems admin
- Queries over dConf, automated actions to perform massive changes, rollbacks

3. Clustrx resource manager

- Modular resource management for clusters
- Open API for external plug-ins
- POSIX-compliant
- Tight integration with Clustrx Watch & dConf
- Adaptive power management for green-computing features to any hardware
- Topology-based resource allocation

Clustrx resource manager (2)



4. Infrastructure management

- Network boot infrastructure for disk & diskless nodes
- Supports all major OSes including Win HPC 2008
- Linearly scalable
- Administrative interface to create and manage bootable OS images
- Accounting of network loads and usage of images

Clustrx green

- User-defined rule-based policies for power management
- Decreasing frequency on computing nodes
- Hibernating on demand
- Powering on and off on demand
- Automated power-down of idle hardware components

Heterogeneous architectures

- Architecture-independent system management
- Hybrid MPI
- Supports accelerated nodes
- Main direction of further development

Clustrx CNL

- Optimized for HPC purposes (97.1% local LINPACK)
- Binary RHEL compatibility
- Supports legacy 32bit applications
- Set of memory managers and CPU schedulers, selectable by resource manage
- Kernel level monitoring and management agent

Our differentiation

- Uniformed, holistic approach
- HPC-grade kernel
- A good part of code was developed from scratch to address Petaflop-scale challenges
- Scalable to 10000s and more nodes
- Cross-platform management

Future directions

- Grid-specific features (tested with CERN`s gLite 3.x,…)
- Extended billing, Monitoring and billing reports, export statistical data
- Virtual clusters
- Adaptive task management based on real-time profiling
- GPU and other accelerators virtualization
- Open MPI tuning and optimization
- Cluster segmentation
- Checkpointing, reliable task run
- Transparent node image migration

The Clustrx Operating System

Scalable and Reliable Next Generation Operating System
for
Petaflop and Exaflop computing

T-PLATFORMS

www.t-platforms.ru