



Clustrx: A new generation operating system designed for HPC

HPC 2010, June 21-25, Cetraro
Dmitry Tkachev, R&D Director

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?

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

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:**

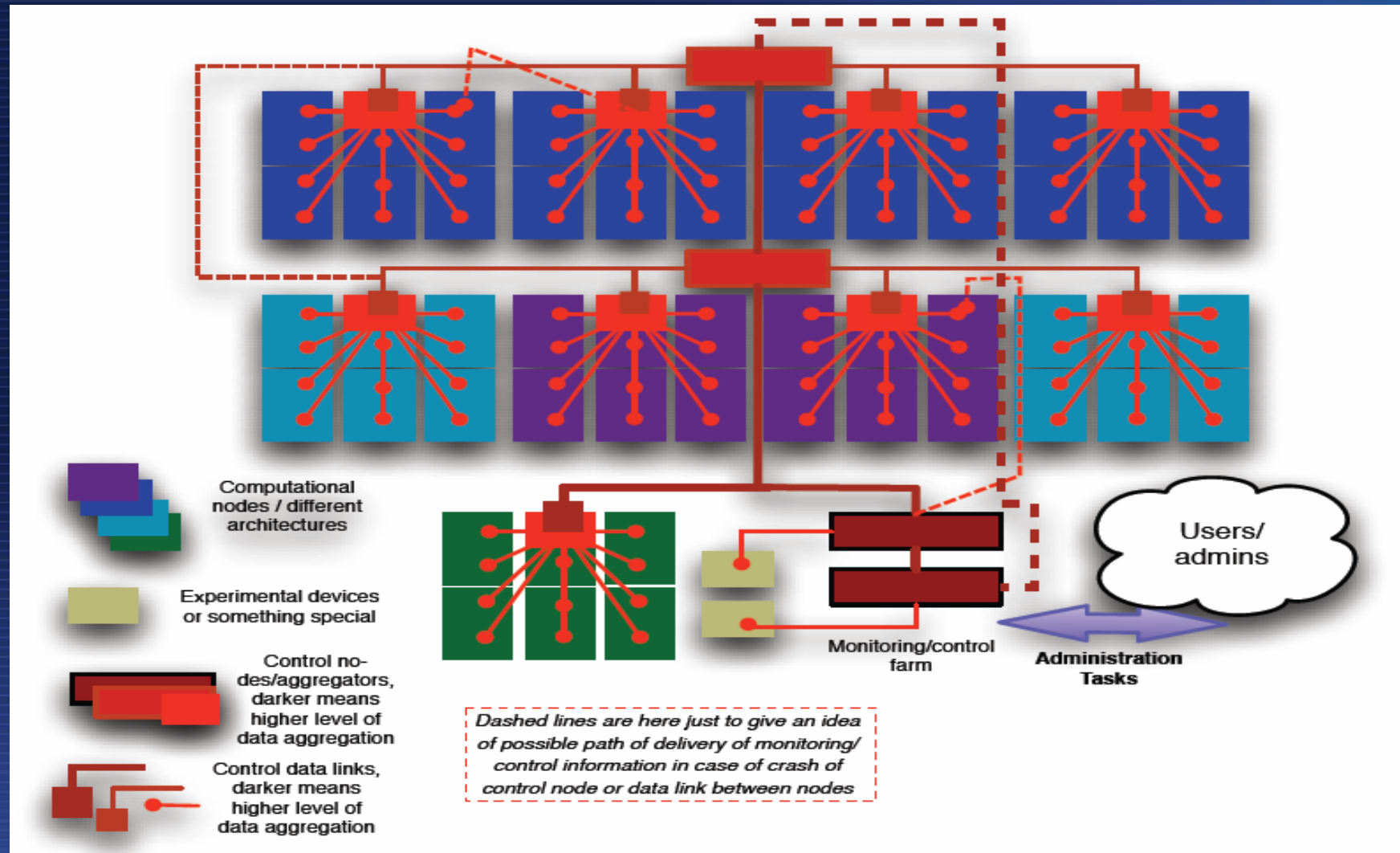
- Emergency switching on/off on particular failures
- Switching off to conserve electricity

- **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



Загружено

Не загружено

Неизвестно

14.03
5:12:00

v 0.3.9



Ряд А

Шкаф R1

Шкаф R2

Шкаф R3

Шкаф R4

Шкаф R5

Шкаф R6

Шкаф R7

Шкаф RE

Шкаф R9

Шкаф R10

Шкаф R11

Меню конфигурирования элементов кластера Scale	
Blade-шасси 06	
Infiniband Leaf 05	
Blade-шасси 05	
Infiniband Leaf 04	
Blade-шасси 04	
Infiniband Leaf 03	
Blade-шасси 03	
Infiniband Leaf 02	
Blade-шасси 02	
Infiniband Leaf 01	
Blade-шасси 01	
CXD ReadyStorage ActiveScale Cluster Part-0 Topology View	

М-ты мониторинга кластера	
Бладе-шасси 12	Infoband Leaf 10
Бладе-шасси 11	Infoband Leaf 09
Бладе-шасси 10	Infoband Leaf 08
Бладе-шасси 09	Infoband Leaf 07
Бладе-шасси 08	Infoband Leaf 06
Бладе-шасси 07	
CXD ReadyStorage	
ActiveScale Cluster Pan-0	
серверы ->	

Компьютер Infineon Core 01
М.п. мониторинга химии защиты «ГП» Вычислительный узел Infineon Leaf 13
Вычислительный узел Infineon Leaf 14
Blade-шасси 16
Infineon Leaf 13
Blade-шасси 15
Infineon Leaf 12
Blade-шасси 14
Infineon Leaf 11
Blade-шасси 13
CXD ReadyStorage ActiveScale ClusterPan-0 архив 1.1.12

Коммутатор
InfiniBand
Core 02

М-ль мониторинга климата
У-волнения воздуха

Коммутатор
C300
01

Коммутатор S50N 01

Сервер резервного
копирования

Ленточная
Библиотека
Quantum
Scalar i500

[illegible]

A vertical rectangular panel with four circular elements arranged vertically. The top and bottom elements are circles with concentric inner circles. The two middle elements are identical circles with concentric inner circles. A horizontal slot is located in the center of the panel, between the two middle circular elements.

M-leaf	monoterpene s derivative acylation (%)
	Blade-waxoil 26
	Infratend Leaf 23
	Blade-waxoil 28
	Infratend Leaf 22
	Blade-waxoil 24
	Infratend Leaf 21
	Blade-waxoil 23
	Infratend Leaf 20
	Blade-waxoil 22
	Infratend Leaf 19
	Blade-waxoil 21

CXJ ReadyStorage
ActiveScale Cluster Pan-
elarray <file>

Модель, модель/серия и компания защитника <1u>
Blade-шасси 32
Infiniband Leaf 29
Blade-шасси 31
Infiniband Leaf 28
Blade-шасси 30
Infiniband Leaf 27

Шкаф R12

Шкаф R13

Шкаф R14

Шкаф R15

Шкаф R16

Шкаф R17

Шкаф R18

Шкаф R19

Шкаф R20

Blade-шасси 38
Infiniband Leaf 37
Blade-шасси 37
Infiniband Leaf 32
Blade-шасси 35
Infiniband Leaf 31
Blade-шасси 35
Infiniband Leaf 30
Blade-шасси 34
Infiniband Leaf 29
Blade-шасси 33
CX4 ReadyStorage ActiveScale Cluster Pan- згруппа <10>

Blade-шасси 44	Infiniband Leaf 38
Blade-шасси 43	Infiniband Leaf 37
Blade-шасси 42	Infiniband Leaf 36
Blade-шасси 41	Infiniband Leaf 35
Blade-шасси 40	Infiniband Leaf 34
Blade-шасси 39	

CX4 ReadyStorage
ActiveScale Cluster Pan-
зума 1x10

Компьютер Infiniband Core 04	
М-ль мониторинга климат датчика "10- заступа" "10-"	
Управляющий узел T80-1	
Управляющий узел T80-2	
KVM-консоль Infiniband Leaf 41	
Blade-шасси 47	
Infiniband Leaf 40	
Blade-шасси 46	
Infiniband Leaf 39	
Blade-шасси 45	
СХД ReadyStorage ActiveScale Cluster Pan-1 серверы 48	

Коммутатор Infiniband Core 05
заглушка <1u> Коммутатор S2410
Коммутатор C300 02
Коммутатор S50N 02 Ус.водопдачи воздуха
заглушка <1u> X 16

Компьютер Infineon Core 08
Мат. мониторная карта Компьютер D-Link D550 Компьютер A/C ISA Manager all cards <10> загрузки <10> загрузки <10>
Blade-шасси 51
Infineon Leaf 44
Blade-шасси 50
Infineon Leaf 43
Blade-шасси 40
Infineon Leaf 42
Blade-шасси 48
CXD ReadyStorage ActiveScale Cluster Par загрузки <10>

A vertical strip of four circles, with the middle one replaced by a rectangle.

М.П. «Информационная система «Ис»	
Blade-шасси 57	Infiniband Leaf 40
Blade-шасси 56	Infiniband Leaf 48
Blade-шасси 55	Infiniband Leaf 47
Blade-шасси 54	Infiniband Leaf 46
Blade-шасси 53	Infiniband Leaf 45
Blade-шасси 52	

CXLD ReadyStorage
ActiveScale Cluster Partition 3/20

207000020

коммутатор
Infiniband
Core 03

М-ль мониторинга климата
заглушка <1u>

1U выч. Узел

1U выч. Узел

1U выч. Узел

1U выч. Узел

Infiniband Leaf 18

Blade-шасси 20

Infiniband Leaf 17

Blade-шасси 19

Infiniband Leaf 16

Blade-шасси 18

Infiniband Leaf 15

Blade-шасси 17

CXD ReadyStorage
ActiveScale Cluster Pan-0

Ряд В

Clustrx Watch (3)

Interface Query Filters

Critical (0) Danger (102) Warning (863) Error (518) Calm (2)

Information Alarms

Rack 6 -> UPS 1

Set ignore flag Power

UPS 1 6.00.01.50

UPS 7 6.00.07.50

Switch 13 6.00.13.55

Switch 27 6.00.27.55

Switch 28 6.00.28.55

Node 39 6.00.39.10

Chassis 14 6.14.00.15

Chassis 15 6.15.00.15

Chassis 16 6.16.00.15

Chassis 17 6.17.00.15

Chassis 18 6.18.00.15

Chassis 19 6.19.00.15

Chassis 21 6.21.00.15

Chassis 22 6.22.00.15

Chassis 23 6.23.00.15

Chassis 24 6.24.00.15

Chassis 25 6.25.00.15

Chassis 26 6.26.00.15

Page 1 of 1 1 - 24 of 24

Search IP Host

System alarms

Critical (0) Danger (102) Error (518) Warning (863) Calm (2)

Last update: 2009-11-11T11:15:21+02:00 Stop update

Key	Alarm	Problem	Hostname	Value	Expiration	Elapsed	Object
2	danger	(node,node_status)	n3048	Not available	29m 40s	6d 15h 34m 31s	Rack 7 -> Node 21
93	danger	(chassis,temp2_status)	Unavailable	danger	21m 42s	4d 22h 39m 32s	Rack 93 -> Chassis 15
2	danger	(node,node_status)	n3036	Not available	29m 55s	6d 15h 34m 31s	Rack 8 -> Node 26
2	danger	(node,node_status)	n3068	Not available	29m 50s	6d 15h 34m 31s	Rack 7 -> Node 42
2	danger	(node,node_status)	n3009	Not available	29m 57s	6d 15h 34m 31s	Rack 5 -> Node 21
2	danger	(node,node_status)	n3015	Not available	29m 58s	6d 15h 34m 31s	Rack 5 -> Node 27
2	danger	(node,node_status)	n3062	Not available	29m 49s	6d 15h 34m 31s	Rack 7 -> Node 36
2	danger	(node,node_status)	n3040	Not available	29m 47s	6d 15h 34m 31s	Rack 7 -> Node 13
2	danger	(node,node_status)	n3066	Not available	29m 59s	6d 15h 34m 31s	Rack 7 -> Node 40
2	danger	(node,node_status)	n3002	Not available	29m 57s	6d 15h 34m 31s	Rack 5 -> Node 14
1	danger	(sensor_status,"MON_CPU_FAN",2)	n3118	13740.0	29m 38s	1m 46s	Rack 6 -> Chassis 23 -> Node 2
2	danger	(node,node_status)	n3021	Not available	29m 46s	6d 15h 34m 31s	Rack 5 -> Node 34
2	danger	(node,node_status)	n3016	Not available	29m 56s	6d 15h 34m 31s	Rack 5 -> Node 28
1	danger	(sensor_status,"MON_V_5VSB",1)	n3134	5.43	29m 41s	29s	Rack 6 -> Chassis 45 -> Node 2
2	danger	(node,node_status)	n3046	Not available	29m 48s	6d 15h 34m 31s	Rack 7 -> Node 19
2	danger	(node,node_status)	n3045	Not available	29m 57s	6d 15h 34m 31s	Rack 7 -> Node 18
2	danger	(node,node_status)	n3018	Not available	29m 47s	6d 15h 34m 31s	Rack 5 -> Node 30
2	danger	(node,node_status)	n3150	Not available	29m 55s	6d 15h 34m 31s	Rack 8 -> Chassis 20 -> Node 2
2	danger	(node,node_status)	n3069	Not available	29m 51s	6d 15h 34m 31s	Rack 7 -> Node 43

Information Alarms Sensors Charts Tasks

Category: All Period: Hourly View type: Area Show levelsLast update: >>

power: advBatteryActualVoltage

AVERAGE: 218.050 MINIMUM: 216.000 MAXIMUM: 219.000

1 - 8 of 8

Search IP Hostname Location location filter1,filter2

[2009-10-01 17:25:00] ERROR Some sh*t happens!

www.i-platforms.ru

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)

The screenshot displays the Clustrx resource manager interface, which is a web-based application for managing tasks and resources. The interface is divided into several sections:

- Tasks queue - ClustrX ...**: This section shows a list of tasks in a table. The table has columns for Job id, Partition, Script, User, State, Time limit, and Reason. The tasks listed are:

Job id	Partition	Script	User	State	Time limit	Reason
27	debug	/home/users/user11/run sleep 100	user11	COMPLETED	600	
28	debug	/home/users/user11/run sleep 100	user11	COMPLETED	600	
29	debug	/home/users/user11/run sleep 100	user11	COMPLETED	600	

- Job info**: This section provides details about the selected job. It includes fields for Job name, State, User name, Job group name, Priority, and Job comment.
- Task configuration**: This section allows users to configure a new task. It includes fields for Task name, Script, Arguments, Working directory, Stdout, Stderr, Operating system, and Maximum restarts allowed. The task name is "my task", the script is "/home/radmin/sleep600.sh", and the arguments are "-a -b -c".
- Prologs/Epilogs**: This section allows users to define prologs and epilogs for the task. It includes fields for Job prolog, Task prolog, Job epilog, and Task epilog.
- Dependencies**: This section allows users to define dependencies for the task. It includes a table with columns for Task and Type.
- Environment**: This section allows users to define environment variables for the task. It includes a table with columns for Variable and Value.
- Notifications**: This section allows users to define notifications for the task. It includes a table with columns for Description, Event, and Handler.

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

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 manager
- Kernel level monitoring and management

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*