Ganglia - an open source monitoring tool

Monitoring of Power Systems – Best Practices

Dr. Michael Perzl (mperzl@de.ibm.com)
IBM Power Systems
Consulting IT Specialist

Good Morning

About me (Michael Perzl):

- Joined IBM in 2000
- Previous job in research and academia
- Working for IBM Germany in Power Systems brand since 2000
  - Currently working for IBM Migration Factory
- Focus areas:
  - AIX
  - Open Source
  - Linux on Power

"Pet Projects":

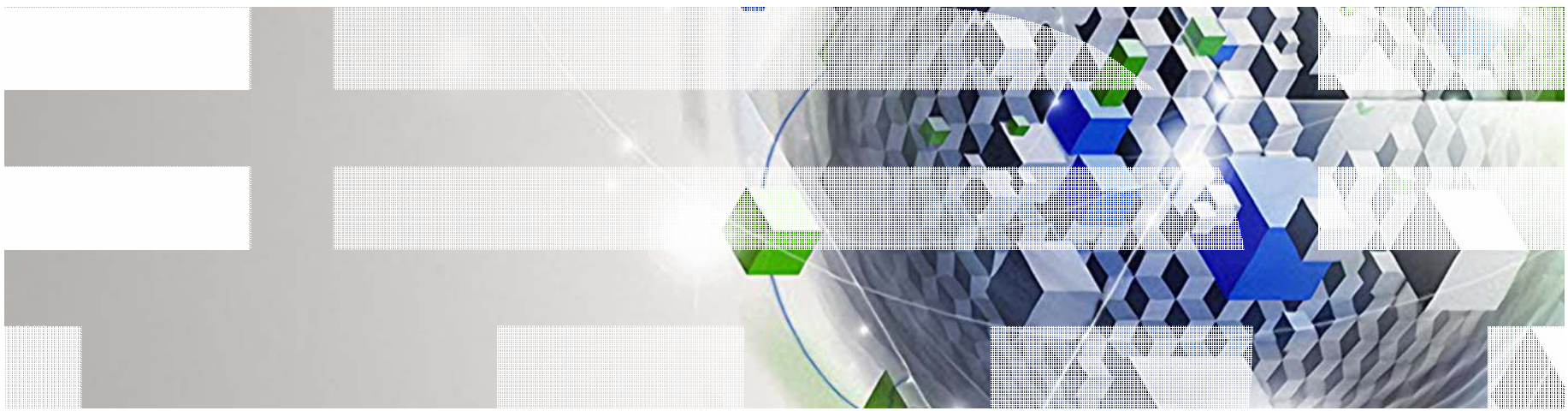
- Ganglia (→http://www.perzl.org/ganglia)
- Large Open Source Repository for AIX (→http://www.perzl.org/aix)

Agenda

- Ganglia – What is it?
- Ganglia Components and Data Flow
- Ganglia Standard Metrics – What can be Monitored?
- Additional Metrics for AIX & Linux on IBM Power Systems
- Ganglia Setup Considerations
- Demo
- Links

- **Please note:**
  - This is **not an IBM product**
  - It is **not officially supported by IBM**

Ganglia – What is it?



Ganglia – What is it? (1/2)

Ganglia properties:

- scalable distributed **monitoring** system for high-performance computing systems such as clusters and grids
- based on a hierarchical design targeted at federations of clusters
- leverages widely used technologies such as
  - XML for data representation
  - XDR (e**X**ternal **D**ata **R**epresentation) for compact, portable data transport
  - Open Source tool **RRDtool** for data storage and visualization
- uses carefully engineered data structures and algorithms to achieve very low per-node overheads and high concurrency
- robust implementation
- BSD-licensed open-source project (written in C) that grew out of the **University of California, Berkeley Millennium Project**

Ganglia – What is it? (2/2)

Ganglia properties (cont.):

- has been ported to an extensive set of operating systems and processor architectures:
  - AIX
  - Darwin
  - FreeBSD
  - HP-UX
  - IRIX
  - Linux
  - OSF
  - NetBSD
  - Solaris
  - Windows (via Cygwin)
- is currently in use on thousands of clusters around the world
- has been used to link clusters across university campuses and around the world
- can scale to handle clusters with 2000+ nodes
  - check http://ganglia.info/ for more details

Demos

- **Wikipedia (check it out!)**
  - The server of the Wikimedia Foundation are monitored with Ganglia and this is made publically available.
- **UC Berkeley Millennium Demo**
  - The UC Berkeley Millennium Project is the birthplace of ganglia. The Millennium Project, which begain in 1998, deployed a hierarchical campus-wide grid of clusters to support advanced scientific computing across dozens of university departments.
- **Grids and Clusters Group Demo**
  - The Grids and Clusters Group at the San Diego Supercomputer Center started bundling ganglia monitoring into their Rocks Installation Tool very early. Years before ganglia was popular, they were submitting patches to the Millennium Group and providing invaluable feedback.

Ganglia Components and Data Flow

Ganglia Components

The ganglia system consists of:

- two unique daemons:
  - Ganglia Monitoring Daemon (**gmond**)
    - monitoring daemon, collects the metrics
    - runs on each node
  - Ganglia Meta Daemon (**gmetad**)
    - polls all gmond clients and stores the collected metrics in Round-Robin Databases (RRDs) via RRDTool
- a PHP-based web frontend
- a few other small utility programs
  - gmetric
    - can be used to easily extend Ganglia with additional user-defined metrics
  - gstat
  - Gexec

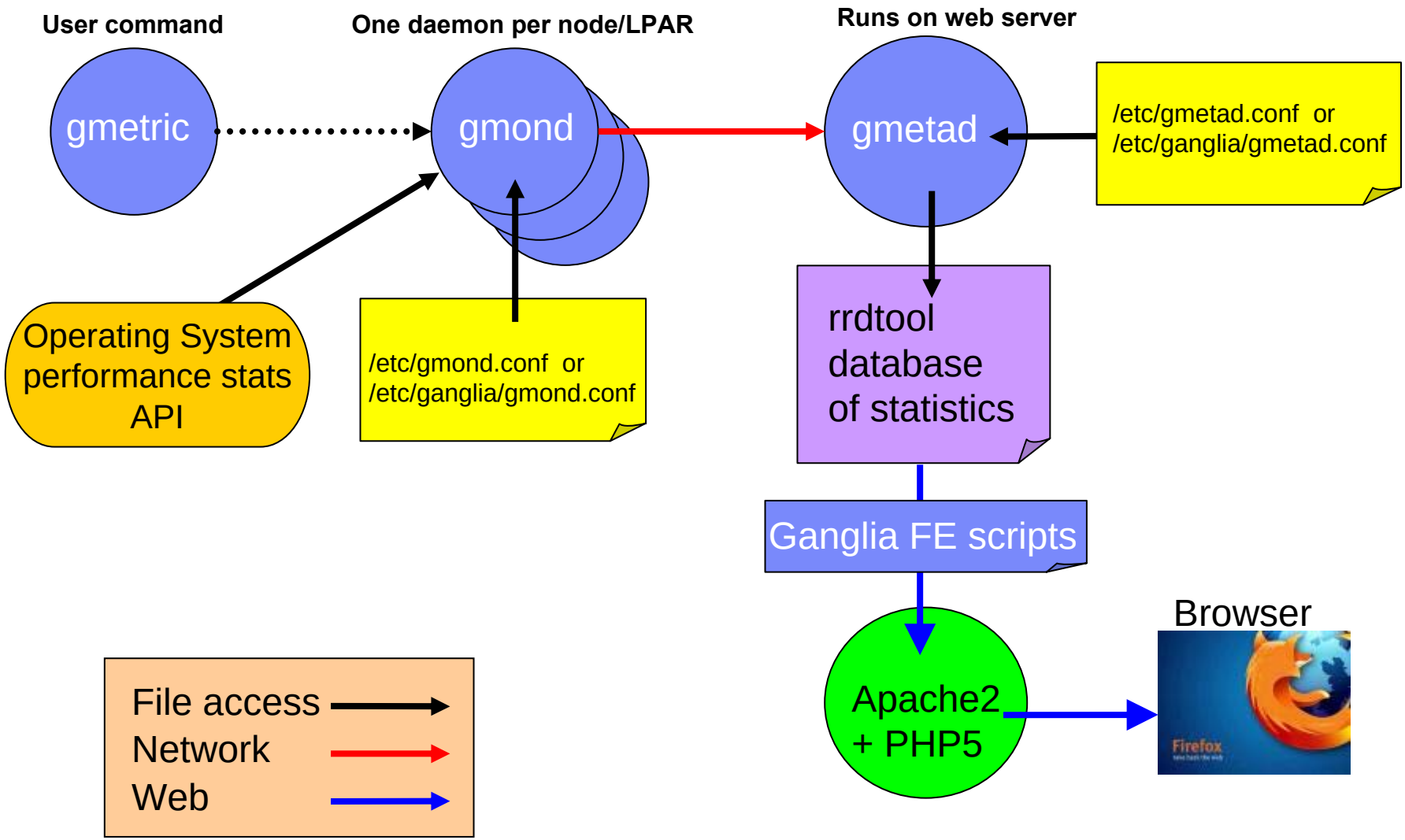
Please note: “Cluster” is used here as a “logical term”!

RRDTool

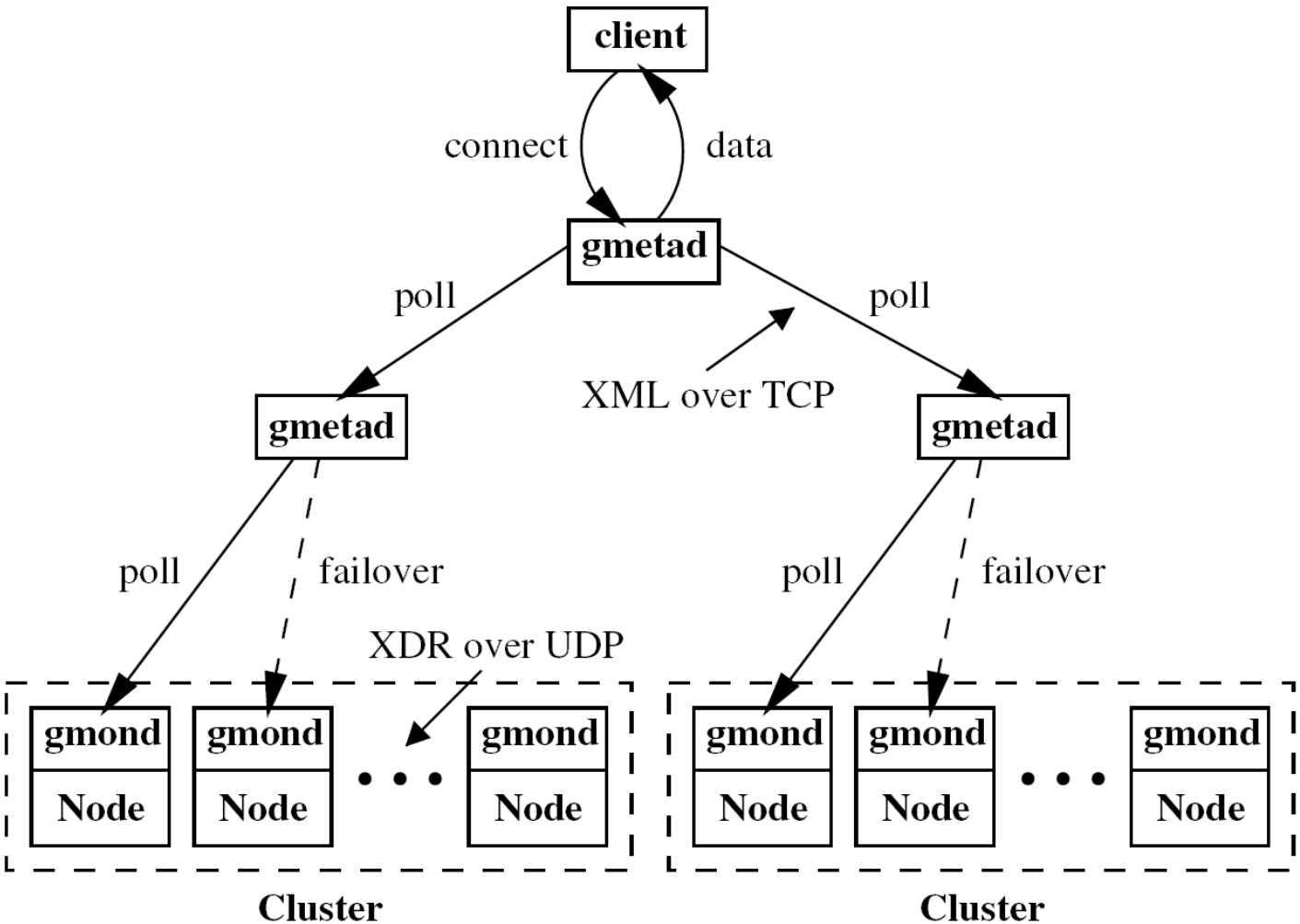
- Homepage: http://oss.oetiker.ch/rrdtool/
- RRD is the Acronym for **R**ound-**R**obin **D**atabase.
- RRD is a system to store and display time-series data (i.e., network bandwidth, machine-room temperature, server load average).
- It stores the data in a very compact way that will not expand over time (**fixed size of DB**), and it presents useful graphs by processing the data to enforce a certain data density.
- It can be used either via simple wrapper scripts (from shell or Perl) or via frontends that poll network devices and put a friendly user interface on it.
- Ganglia uses RRDTool for storing and graphing all data

RRDTool is the industry standard tool to store and display time-series data!

Ganglia – Data Flow



Ganglia Architecture and Communication



Ganglia Standard Metrics – What can be Monitored ?

Power your planet.

Metrics

Definition of a metric:

- A metric is a certain observed property of the system.

Number of metrics:

- 34 standard metrics, i.e., available (i.e., defined) on all platforms
- Additional platform dependent metrics available
  - Solaris
    - 8 additional metrics available
  - HP-UX
    - 4 additional metrics available
  - AIX
    - In default configuration none, details later….

Remarks:

- One RRD database per Ganglia metric is used
- Database size is fixed (ca. 12 kB per RRD database with default settings for gmetad "RRAs" stanza), details later
- Some standard metrics do not exist on all platforms, e.g., some metrics (coming from Linux) don't exist or don't make sense on AIX

Ganglia Standard Metrics

1) boottime
2) bytes_in
3) bytes_out
4) cpu_aidle
5) cpu_idle
6) cpu_nice
7) cpu_num
8) cpu_intr
9) cpu_sintr
10) cpu_speed
11) cpu_system
12) cpu_user
13) cpu_wio
14) disk_free
15) disk_total
16) load_one
17) load_five
18) load_fifteen
19) machine_type
20) mem_total
21) mem_free
22) mem_shared
23) mem_buffers
24) mem_cached
25) mtu
26) os_name
27) os_release
28) part_max_used (Linux specific)
29) pkts_in
30) pkts_out
31) proc_run
32) proc_total
33) swap_free (on AIX: paging space)
34) swap_total (on AIX: paging space)

Additional Metrics for AIX & Linux on IBM Power Systems

Power your planet.

Current Deficiences of Ganglia on Power5/6/7

- Ganglia does not understand Power5/6/7 Shared Processor LPAR statistics
  - things like capped, weight, CPU entitlement etc…
- Ganglia provides no individual Ethernet adapter monitoring
- Ganglia provides no individual Fibre Channel adapter monitoring
- Ganglia provides no individual Disk monitoring
- Ganglia does not understand Power6/7 Active Memory Sharing (AMS) statistics
- Ganglia does not understand Power7 Active Memory Expansion (AME) statistics
- Ganglia provides no IBM rPerf nor SPEC CPU2006 statistics

Adding Metrics to Ganglia

- Easy solution:
  - Extend Ganglia with the utility program **gmetric**
  - Details in appendix "Extending Ganglia with gmetric"

- Preferred solution:
  - Add these new metrics to the gmond implementation on AIX and Linux on Power
    - Requires significant patching of Ganglia source code for Ganglia V3.0.X
  - Starting with Ganglia V3.1.X support for DSO modules (= dynamically loadable extensions) is available
    - Can be built either with C/C++ or Python
    - DSO support available for AIX and Linux on Power
    - Separation of core Ganglia source code possible

Additional available Ganglia DSO Modules

- DSO for IBM Power extensions (module `mod_ibmpower`)
- DSO for IBM rPerf and SPEC CPU2006 metrics (module `mod_ibmrperf`)
- DSO for Active Memory Expansion (AME) (module `mod_ibmame`)
- DSO for Active Memory Sharing (AMS) (module `mod_ibmams`)
- AIX DSO for Fibre Channel devices (module `mod_ibmfc`)
- AIX DSO for Network devices (`mod_ibmnet`)
- Linux DSO for Network devices (`mod_netif`)
- AIX DSO for Hard Disk devices (`mod_aixdisk`)
- Linux DSO for Hard Disk devices (`mod_linuxdisk`)

New

IBM Power Systems DSO Support (Version ≥ 3.1.X) (1/5)

mod_ibmpower:

- The Power5/6/7 extensions (22 metrics) are contained in a separate DSO module (written in C) called "`mod_ibmpower`".
- If installed, this DSO module is loaded during runtime/startup of gmond.
- Config file: `/etc/ganglia/conf.d/ibmpower.conf`

mod_ibmrperf:

- The IBM rPerf and SPEC CPU2006 extensions (5 metrics) are contained in a separate DSO module (written in C) called "`mod_ibmrperf`".
- If installed, this DSO module is loaded during runtime/startup of gmond.
- Config file: `/etc/ganglia/conf.d/ibmrperf.conf`

IBM Power Systems DSO Support (Version ≥ 3.1.X) (2/5)

mod_ibmame:

- The Power7 Active Memory Expansion (AME) extensions (11 metrics) are contained in a separate DSO module (written in C) called "`mod_ibmame`".
- If installed, this DSO module is loaded during runtime/startup of gmond.
- Config file: `/etc/ganglia/conf.d/ibmame.conf`

mod_ibmams:

- The Power6/7 Active Memory Sharing (AMS) extensions (9 metrics) are contained in a separate DSO module (written in C) called "`mod_ibmams`".
- If installed, this DSO module is loaded during runtime/startup of gmond.
- Config file: `/etc/ganglia/conf.d/ibmams.conf`

IBM Power Systems DSO Support (Version ≥ 3.1.X) (3/5)

mod_ibmfc (AIX only):

- The extensions (maximum of 4 metrics per single device) for individual Fibre Channel devices are contained in a separate DSO module (written in C) called "`mod_ibmfc`".
- If installed, this DSO module is loaded during runtime/startup of gmond.
- Config file: `/etc/ganglia/conf.d/ibmame.conf`

mod_ibmnet (AIX only):

- The extensions (maximum of 4 metrics per single device) for individual Ethernet devices are contained in a separate DSO module (written in C) called "`mod_ibmnet`".
- If installed, this DSO module is loaded during runtime/startup of gmond.
- Config file: `/etc/ganglia/conf.d/ibmnet.conf`

IBM Power Systems DSO Support (Version ≥ 3.1.X) (4/5)

mod_netif (Linux only):

- The extensions (maximum of 4 metrics per single device) for individual Ethernet devices are contained in a separate DSO module (written in C) called "`mod_netif`".
- If installed, this DSO module is loaded during runtime/startup of gmond.
- Config file: `/etc/ganglia/conf.d/ibmnet.conf` (Linux)

mod_aixdisk (AIX only):

- The extensions (maximum of 20 metrics per single device) for individual hard disk devices are contained in a separate DSO module (written in C) called "`mod_aixdisk`".
- If installed, this DSO module is loaded during runtime/startup of gmond.
- Config file: `/etc/ganglia/conf.d/aixdisk.conf`

IBM Power Systems DSO Support (Version ≥ 3.1.X) (5/5)

mod_linuxdisk (Linux only):

- The extensions (maximum of 11 metric per single device) for individual hard disk devices are contained in a separate DSO module (written in C) called "`mod_linuxdisk`" (Linux).
- If installed, this DSO module is loaded during runtime/startup of gmond.
- Config file: `/etc/ganglia/conf.d/linuxdisk.conf`

DSO for IBM Power Extensions

Ganglia Power5/6/7 Metrics

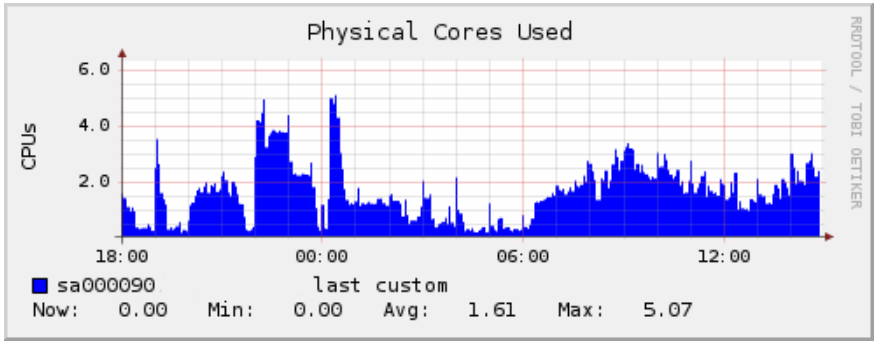
23 additional metrics for AIX & Linux:

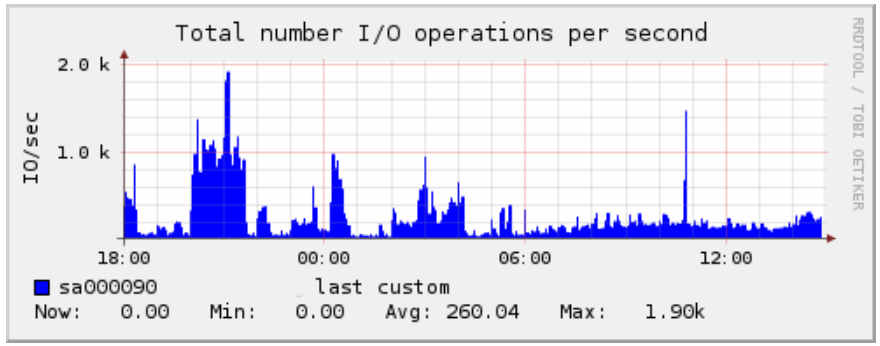
1) capped
2) cpu_entitlement
3) cpu_in_lpar
4) cpu_in_machine
5) cpu_in_pool
6) cpu_pool_id
7) cpu_pool_idle
8) cpu_used
9) disk_read
10) disk_write
11) disk_iops
12) fwversion
13) kernel64bit
14) lpar
15) lpar_name
16) lpar_num
17) modelname
18) oslevel
19) serial_num
20) smt
21) splpar
22) weight

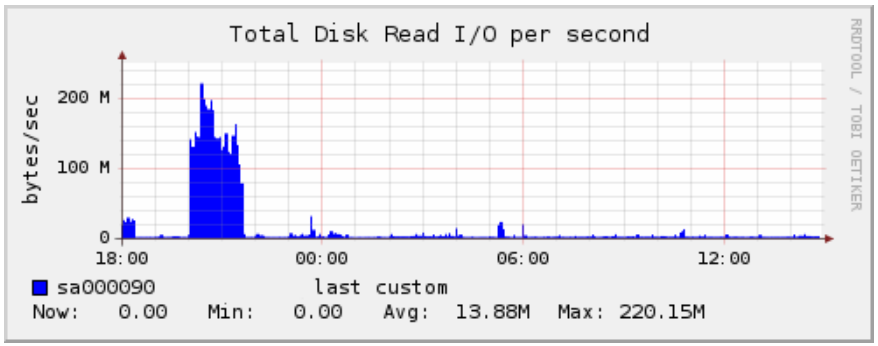
For Power6/7 only (at least AIX V5.3 TL07 required):

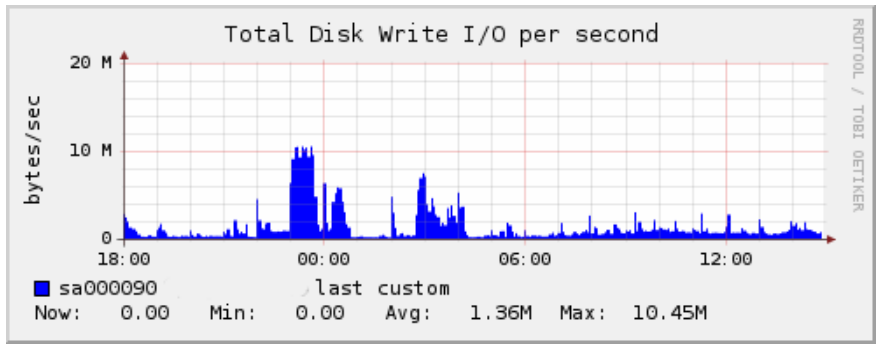
23) cpu_in_syspool (on a Power5 system: same value as cpu_in_pool)

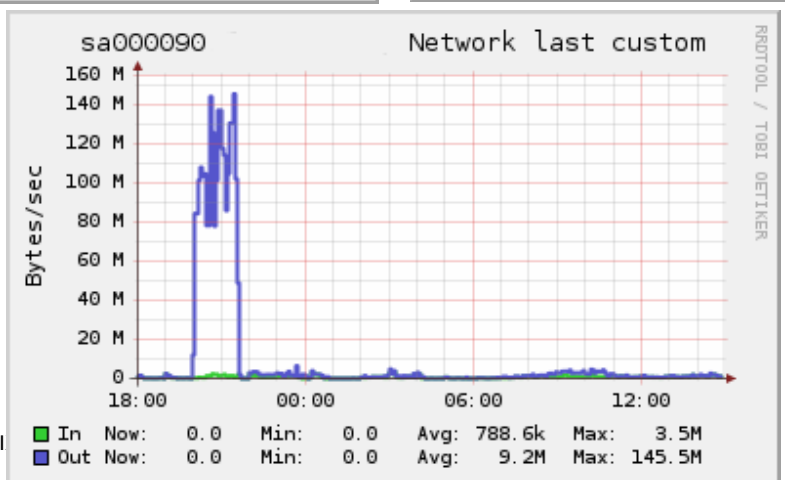
Example AIX LPAR (running SAP + Oracle)



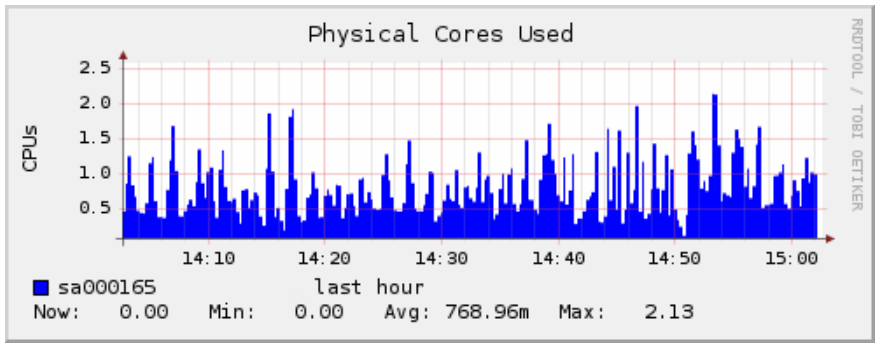


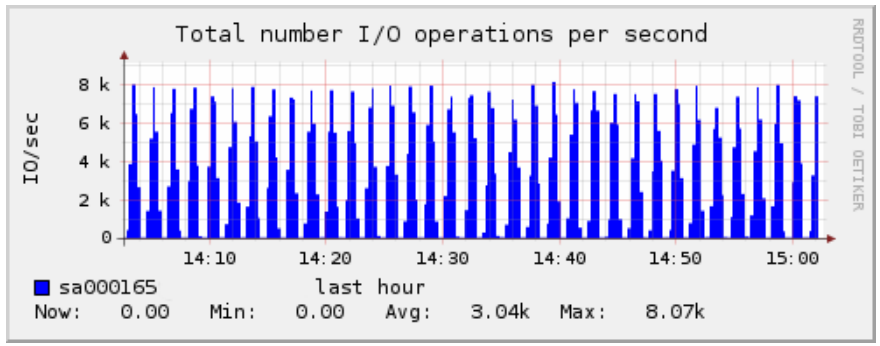


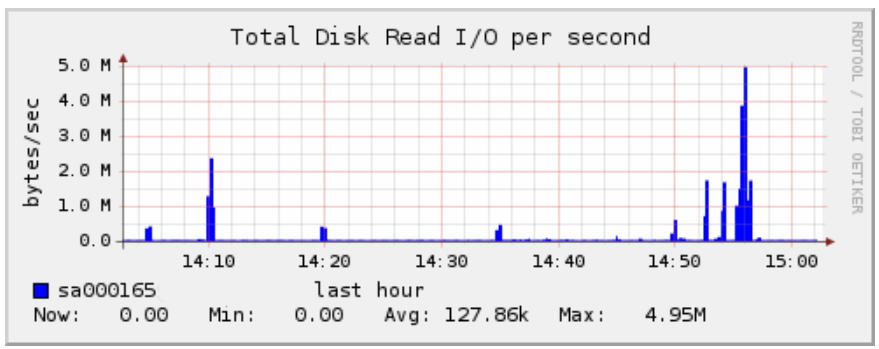


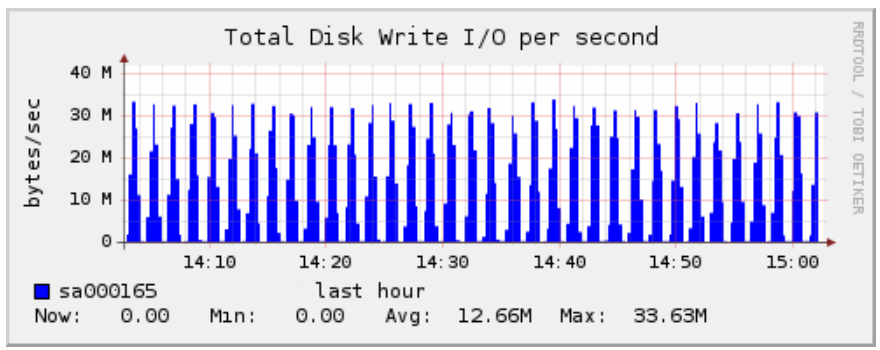


Ganglia gmetad (AIX) for ~560 AIX systems (Power4,5,6,7) Performance Statistics (1/2), last hour view

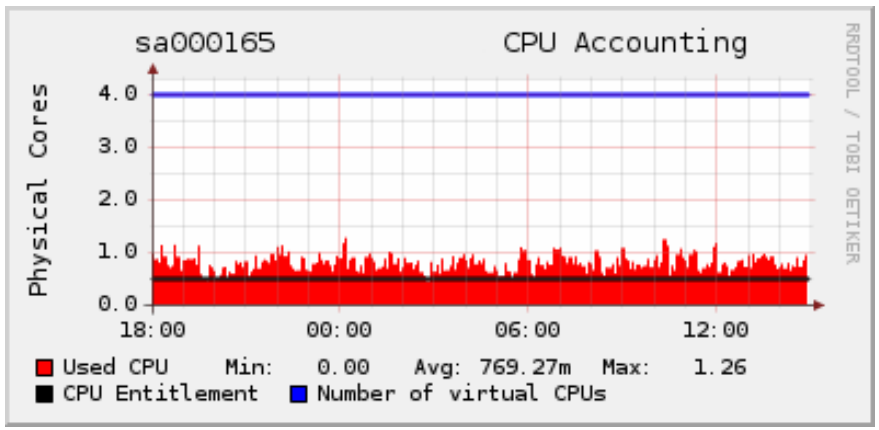


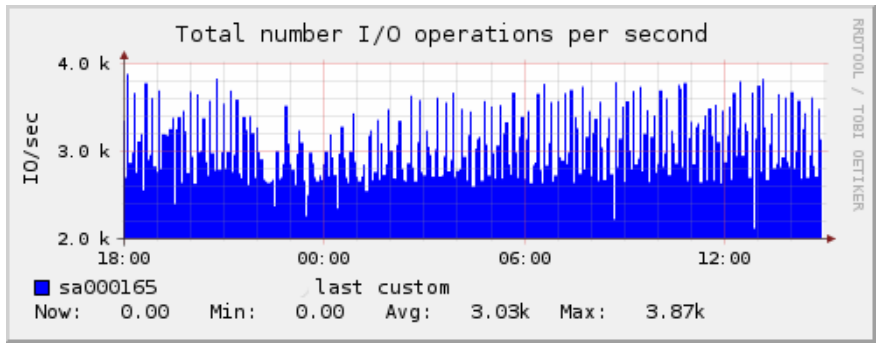


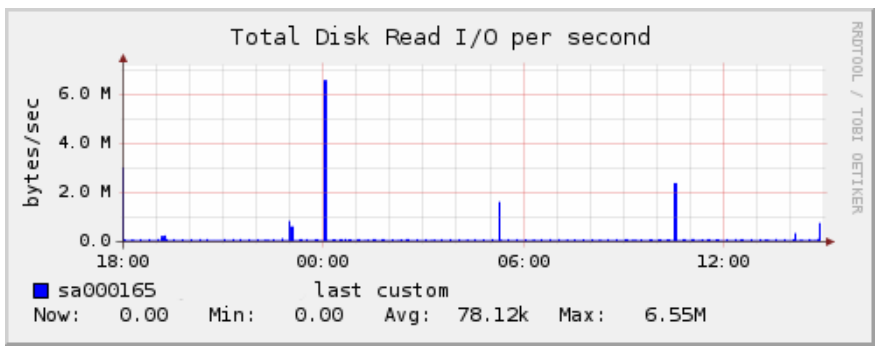


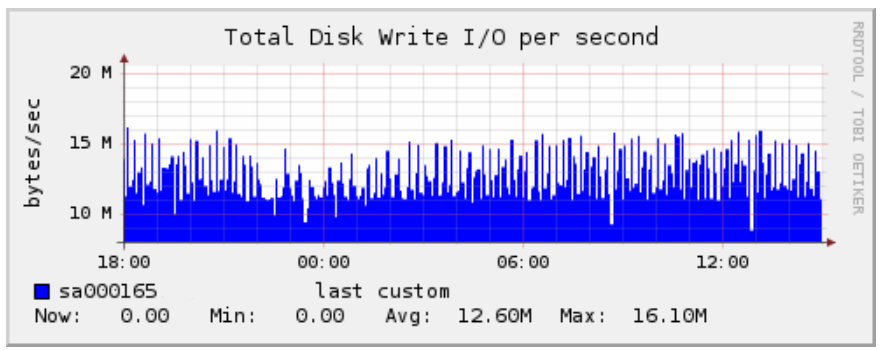


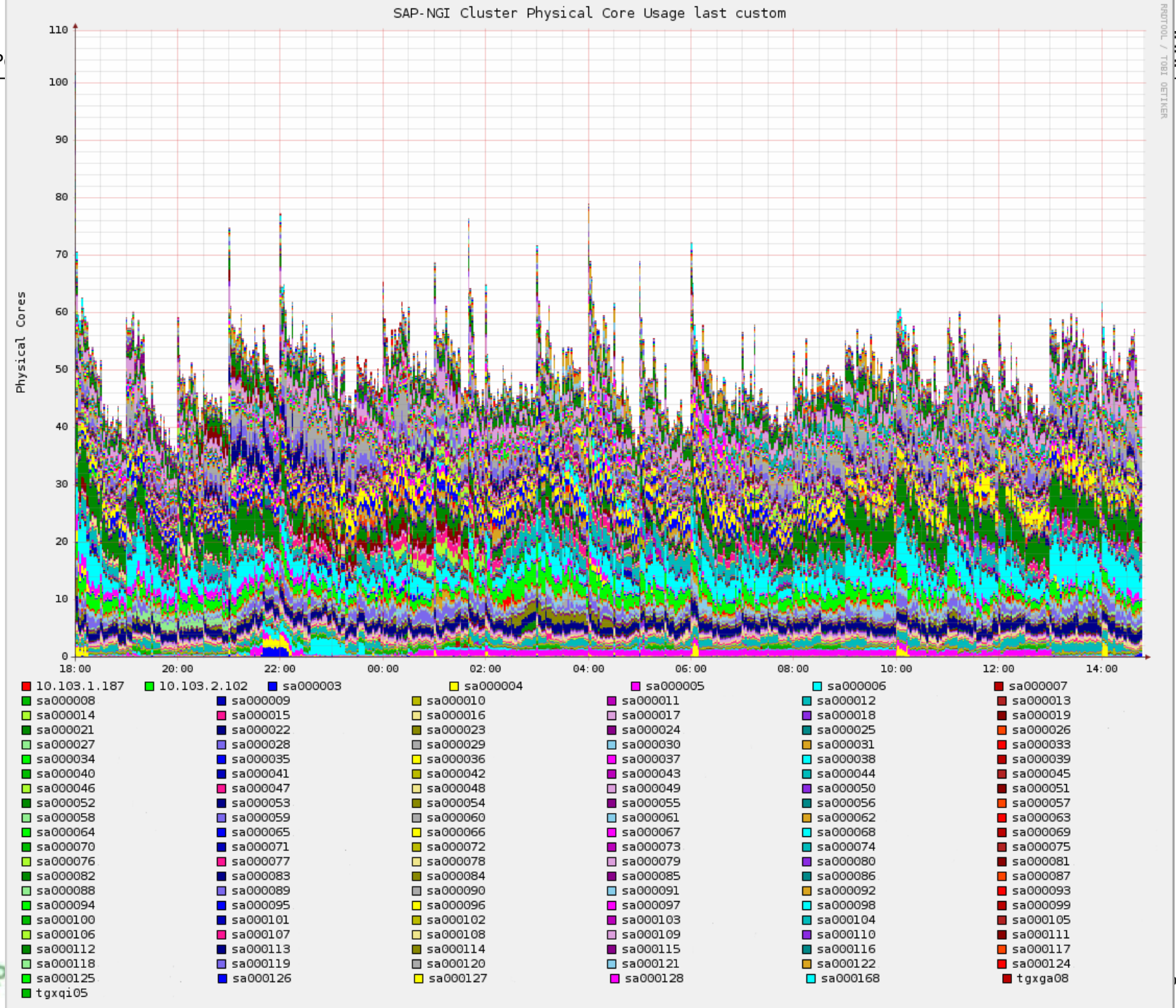
Ganglia gmetad (AIX) for ~560 AIX systems (Power4,5,6,7) Performance Statistics (2/2), custom time interval



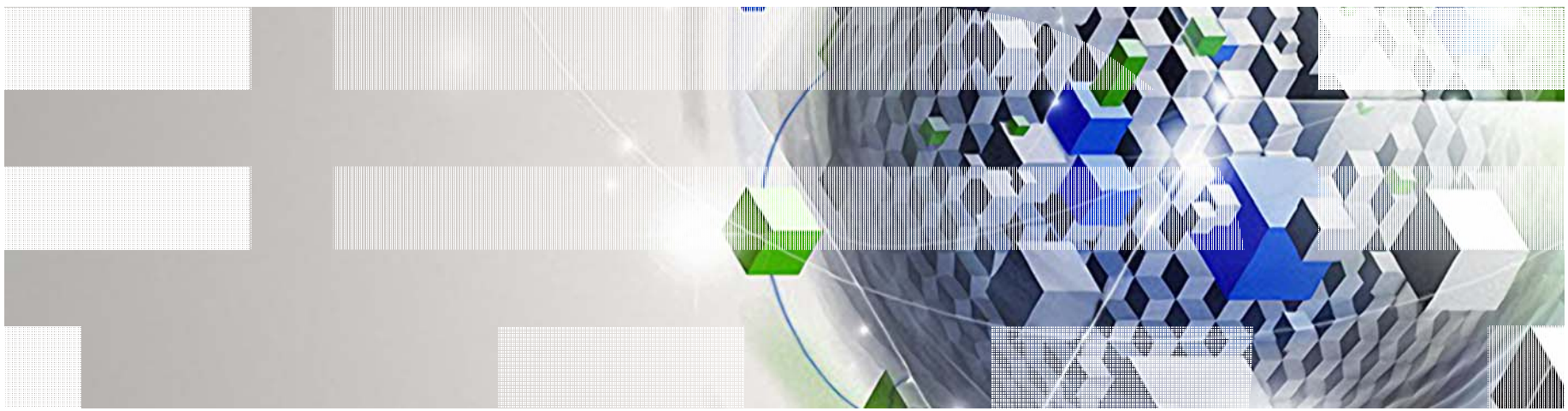




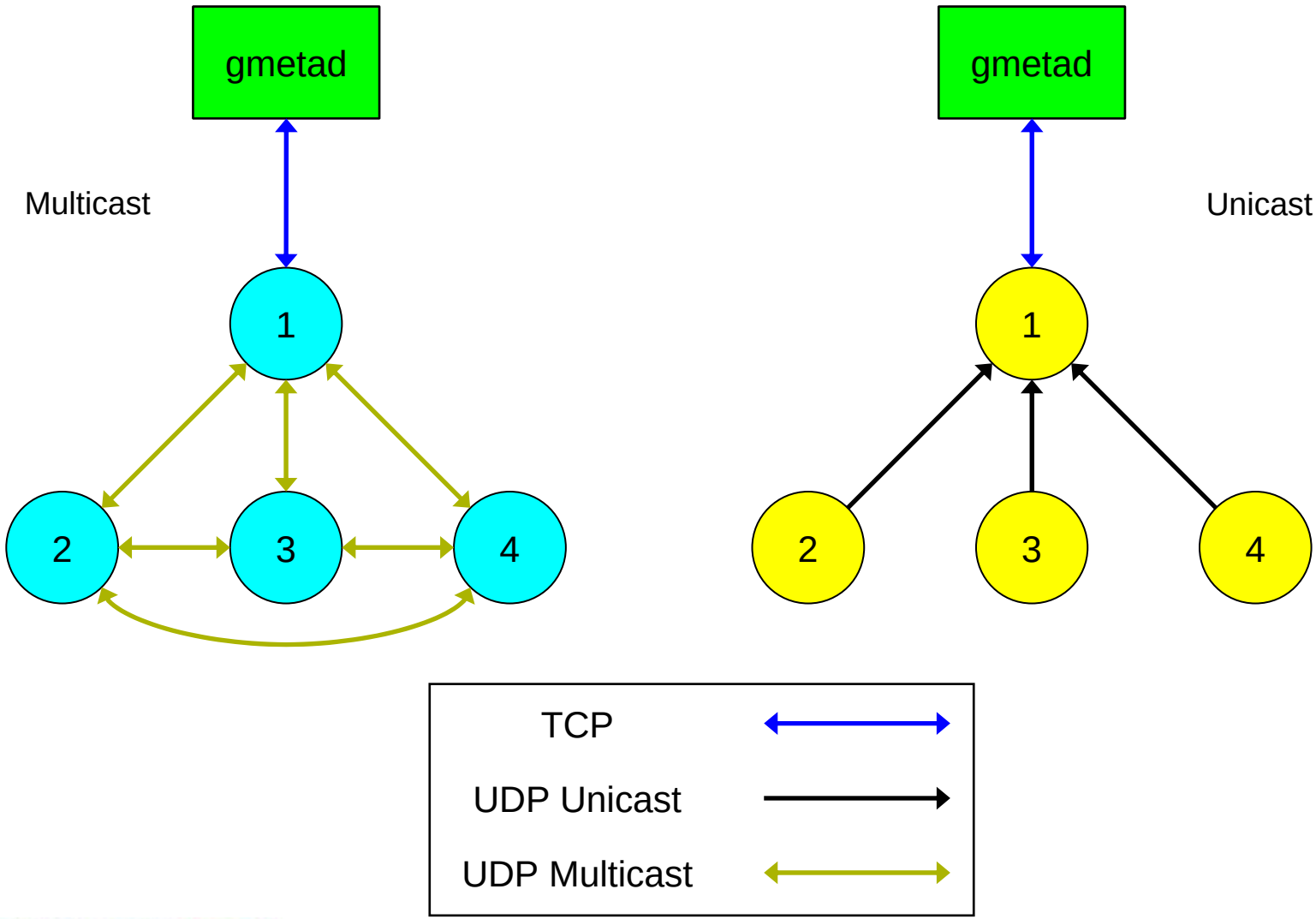



SAP-NGI Cluster Physical Core Usage last custom
Physical Cores
110
100
90
80
70
60
50
40
30
20
10
0
18:00
20:00
22:00
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
10.103.1.187
10.103.2.102
sa000003
sa000004
sa000005
sa000006
sa000007
sa000008
sa000009
sa000010
sa000011
sa000012
sa000013
sa000014
sa000015
sa000016
sa000017
sa000018
sa000019
sa000021
sa000022
sa000023
sa000024
sa000025
sa000026
sa000027
sa000028
sa000029
sa000030
sa000031
sa000033
sa000034
sa000035
sa000036
sa000037
sa000038
sa000039
sa000040
sa000041
sa000042
sa000043
sa000044
sa000045
sa000046
sa000047
sa000048
sa000049
sa000050
sa000051
sa000052
sa000053
sa000054
sa000055
sa000056
sa000057
sa000058
sa000059
sa000060
sa000061
sa000062
sa000063
sa000064
sa000065
sa000066
sa000067
sa000068
sa000069
sa000070
sa000071
sa000072
sa000073
sa000074
sa000075
sa000076
sa000077
sa000078
sa000079
sa000080
sa000081
sa000082
sa000083
sa000084
sa000085
sa000086
sa000087
sa000088
sa000089
sa000090
sa000091
sa000092
sa000093
sa000094
sa000095
sa000096
sa000097
sa000098
sa000099
sa000100
sa000101
sa000102
sa000103
sa000104
sa000105
sa000106
sa000107
sa000108
sa000109
sa000110
sa000111
sa000112
sa000113
sa000114
sa000115
sa000116
sa000117
sa000118
sa000119
sa000120
sa000121
sa000122
sa000124
sa000125
sa000126
sa000127
sa000128
sa000168
tgxga08
tgxqi05

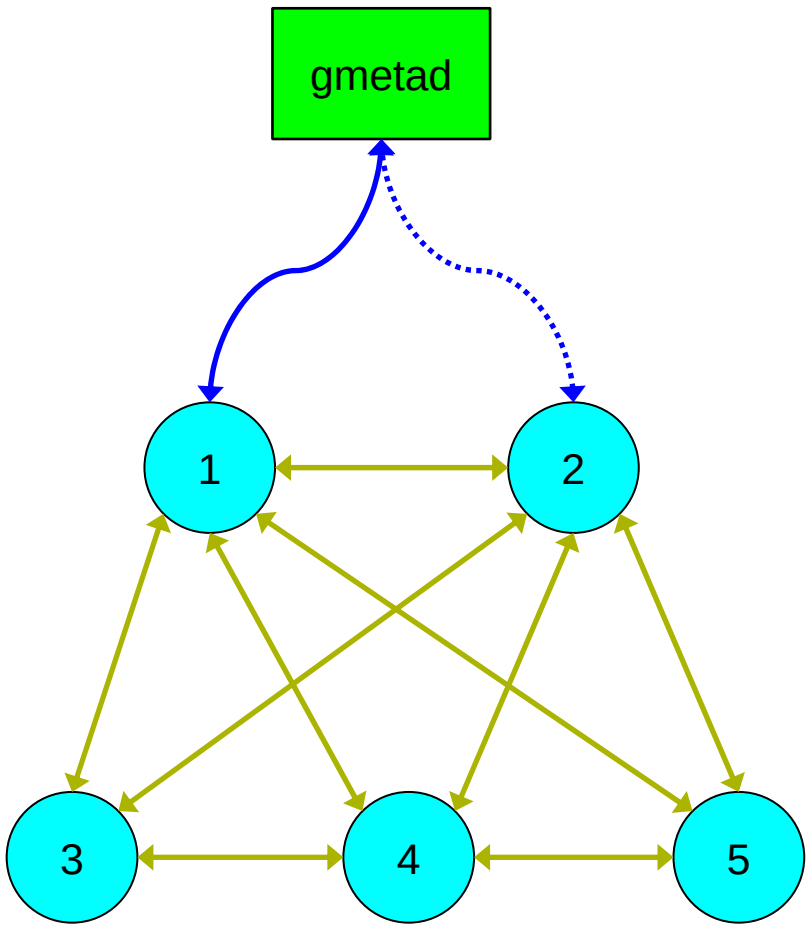
Ganglia Setup Considerations



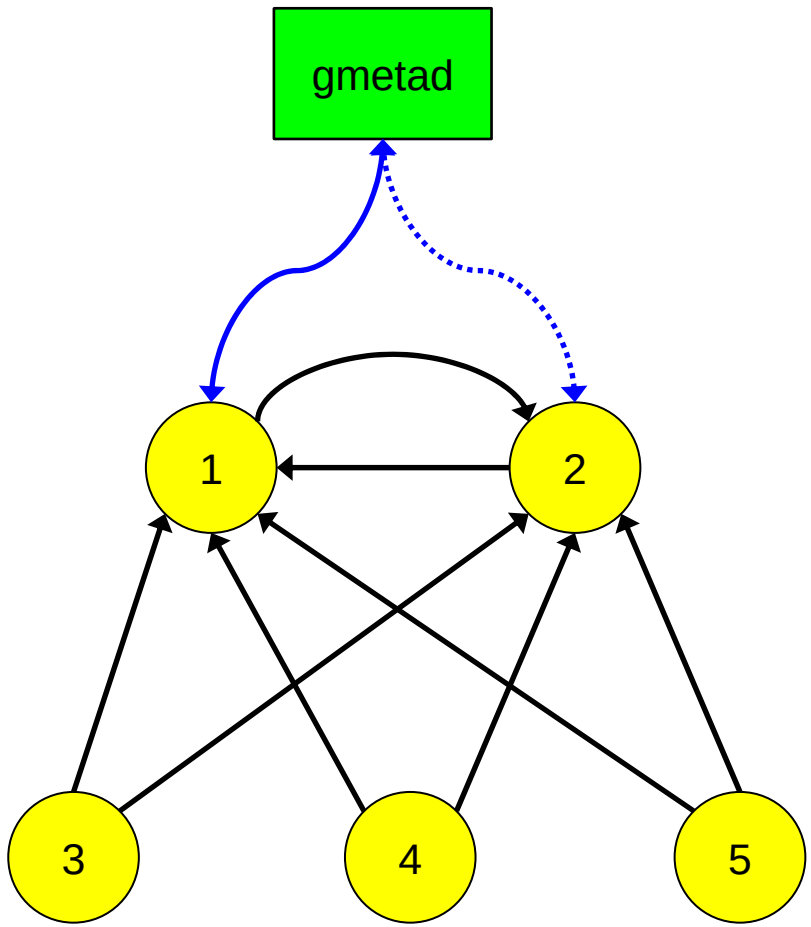
Ganglia Communication: Multicast vs. Unicast



Ganglia Multicast Setup vs. Unicast Setup



Multicast



Unicast

Ganglia Multicast Setup vs. Unicast Setup

Multicast Setup

- Advantages
  - Easy setup, no “sophisticated architecture“ required
- Disadvantages
  - “Everybody knows everything of everybody“ (and doesn‘t forget easily)
  - Setup changes require restart of all gmonds

Unicast Setup

Preferred Setup

- Advantages
  - Exact communication structure must be given
  - Setup changes require much less work compared to multicast setup
- Disadvantages
  - More complex setup, “must think before setup“

Setup Example

Machines considered:

- Dual VIOS Power system, (e.g., p7 770, i.e. LPM capable) gmond
- Single VIOS Power system, (e.g., p7 730, i.e., LPM capable) gmond
- Standalone Power system, (e.g., p4 615, i.e., non LPM capable) gmond

Types of LPARs:

- VIO Server
- DB LPARs
- SAP LPARs
- AppServer LPARs

Comparison of recommended setups:

- before POWER6 and Live Partition Mobility
- now with Live Partition Mobility

Recommended Setup “before“ Live Partition Mobility

Recommended setup was:

- “Cluster“ all LPARs of a physical system together

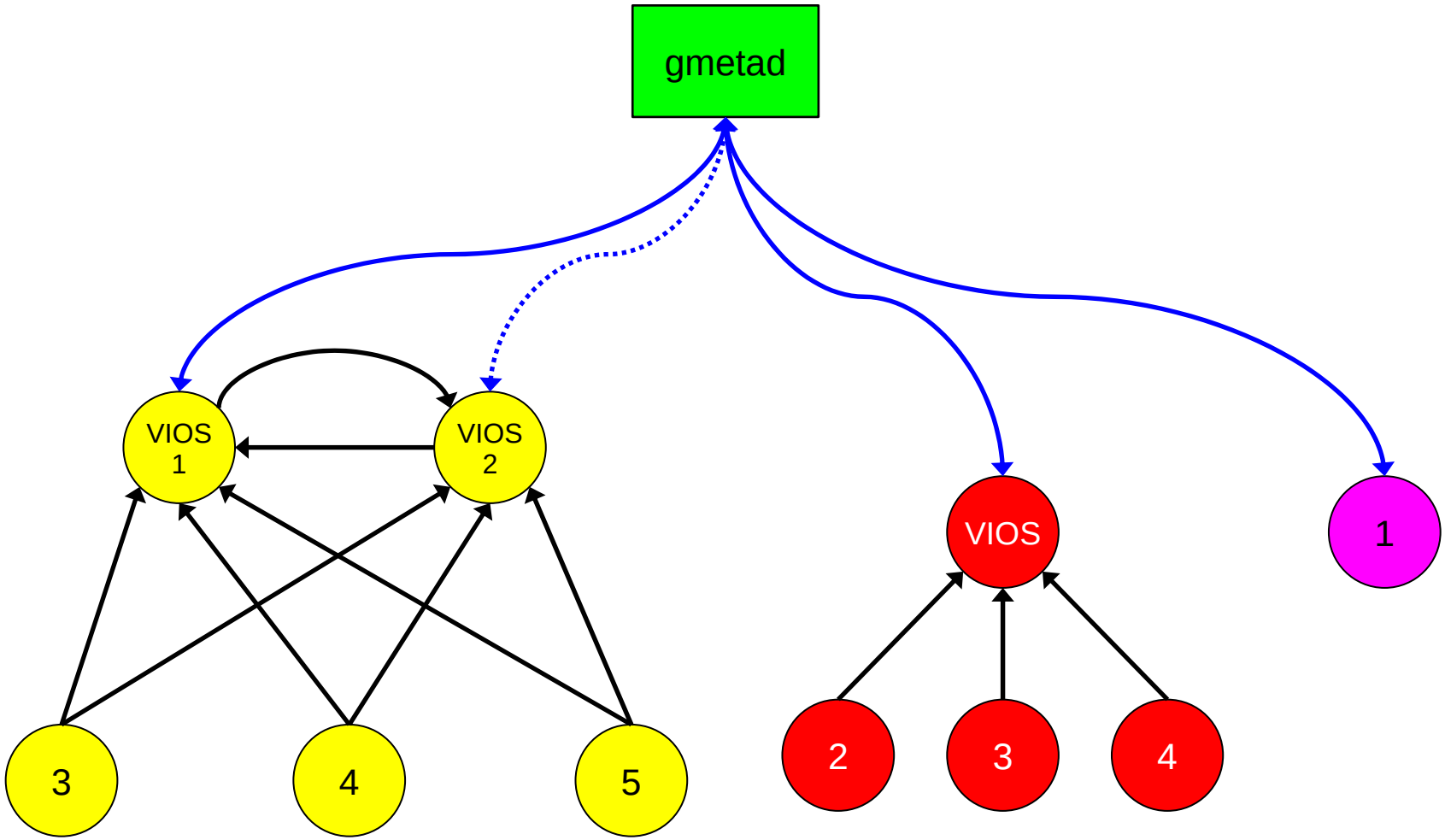
gmond Communication setup:

- Dual VIOS Power system:
  - All LPARs on this box send their data to both VIO Servers on this box
  - Both VIO Servers also exchange their performance information
- Single VIOS Power system:
  - All LPARs on this box send their data to the VIO Server on this box
- Single system:
  - Send nothing

Assumption:

- An LPAR never migrates from a physical box to another one! (true for Power5)

Setup Example “before“ Live Partition Mobility



Live Partition Mobility and its implications

Problem:

- A Live Partition Migration operation moves a LPAR from one physical box to another one
- Previous “hardware-based“ setup not applicable anymore for LPM-capable LPARs!
  - Must notify all involved gmonds/gmetads of migrated LPAR
    → must move stored RRD files to new “cluster location“

Solution:

- “Cluster“ all LPARs logically, i.e., according to their “type“
  - Cluster all VIO Server LPARs together
  - Cluster all DB LPARs together
  - Cluster all SAP LPARs together
  - Cluster all AppServer LPARs together
  - etc.

Recommended Setup “after“ Live Partition Mobility

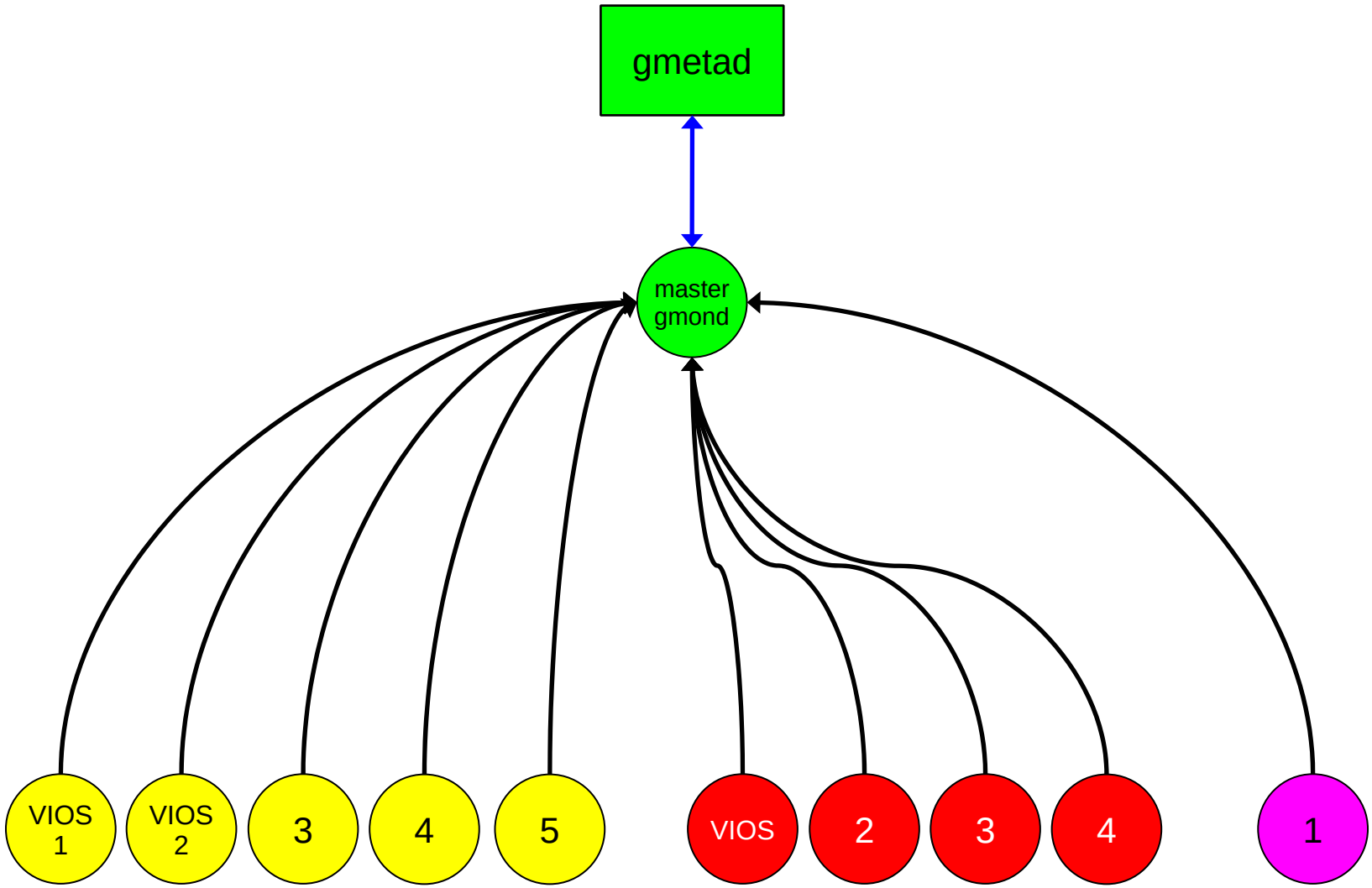
Rationale:

- A SAP LPAR is still a SAP LPAR after a Live Partition Migration!
- A DB LPAR is still a DB LPAR after a Live Partition Migration!
- etc.

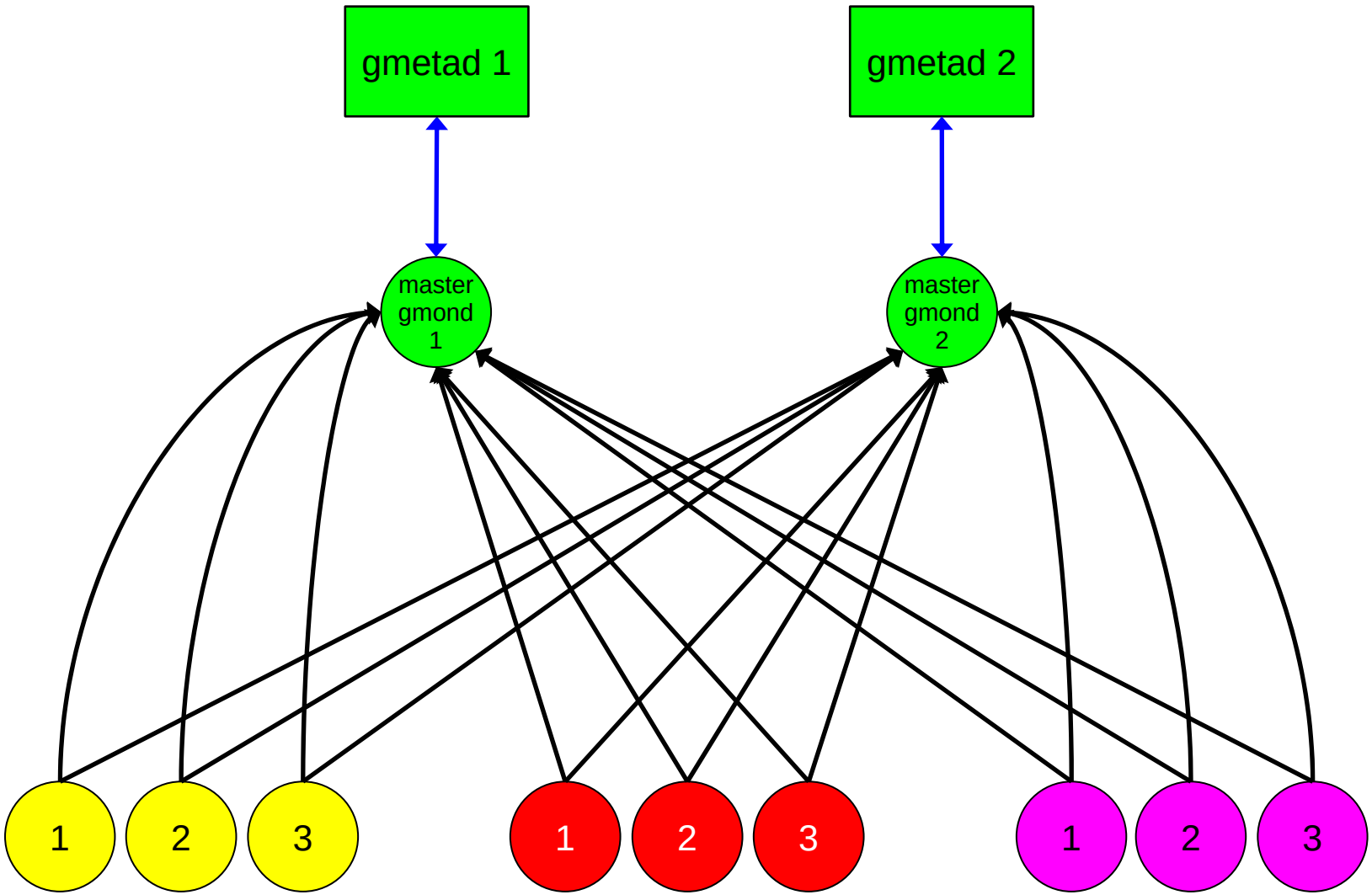
gmond Communication setup:

- Dual VIOS Power system:
  - All LPARs including VIO Servers on this box send their data to the “master gmond“
- Single VIOS Power system:
  - All LPARs including VIO Server on this box send their data to the “master gmond“
- Single system:
  - Send the data to the “master gmond“

Setup Example “after“ Live Partition Mobility



Ganglia Unicast, Multihomed gmonds, “HA-Setup“



“Physical Box View“ still possible?

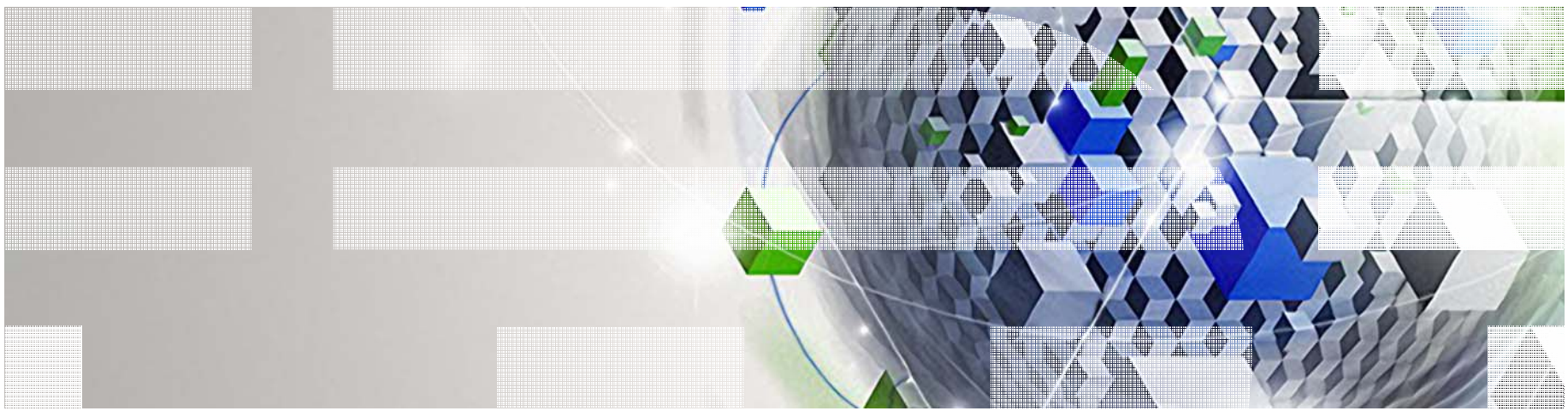
Question:

- How do I get my “physical box view now“?

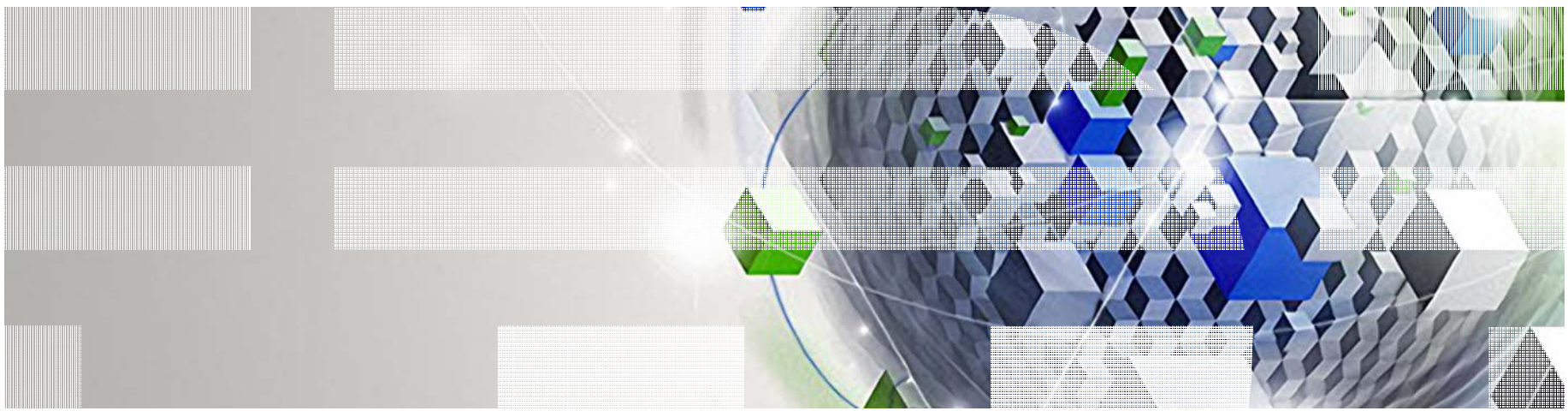
Answer:

- Use the new Web 2.0 GUI interface and define “Views“!

Demo



Links



Links (1/2)

- Main Ganglia website
  - http://ganglia.info/
- Ganglia Documentation
  - http://ganglia.info/docs/
- Ganglia Source Code Download
  - http://ganglia.sourceforge.net/downloads.php
- Ganglia Power5/6/7 extensions and ready-to-run binaries (RPM files) as well as source code
  - http://www.perzl.org/ganglia/
  - http://www.perzl.org/aix/index.php?n=Main.Ganglia
- My personal AIX Open Source repository
  - http://www.perzl.org/aix/

Links (2/2)

- Ganglia Usage at Wikipedia
  - http://ganglia.wikimedia.org/
- RRDTool homepage
  - http://oss.oetiker.ch/rrdtool/
- Ganglia How-To on IBM AIX wiki site
  - http://www.ibm.com/developerworks/wikis/display/WikiPtype/ganglia
- Open Source with AIX on IBM AIX wiki site
  - http://www.ibm.com/developerworks/wikis/display/wikiptype/aixopen
- IBM AIX wiki site:
  - https://www.ibm.com/developerworks/wikis/display/WikiPtype/AIX
- IBM Linux on Power wiki site:
  - https://www.ibm.com/developerworks/wikis/display/LinuxP/Home

Questions ?

Thank you for your attention !