



High-density Grid storage system optimization at ASGC

Shu-Ting Liao
ASGC Operation team
ISGC 2011



Outline

- Introduction to ASGC Grid storage system
- Storage status and issues in 2010
- Storage optimization
- Summary

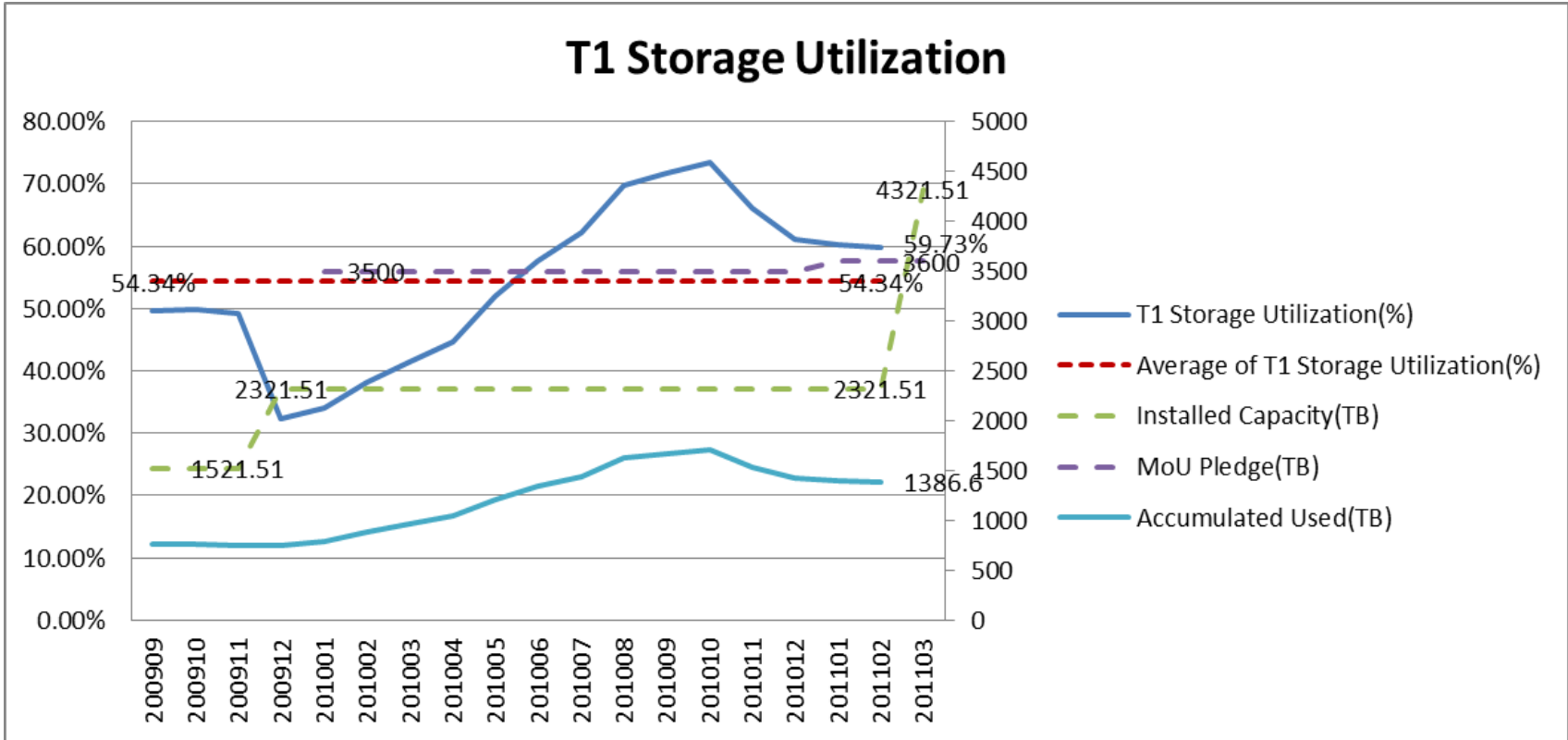


Introduction

- ASGC provides high intensive data services for both WLCG Tier-1 (Taiwan-LCG2) and Tier-2 (TW-FTT) center by **CASTOR** and **DPM**.
- Using more than **120 disk servers** to support **~5.8 PB** of inbound and outbound data for ATLAS and CMS experiments during 2010 data taking.
- Optimizing capacity and storage efficiency in a complex storage architecture and limited space of datacenter is a big challenging task.



T1 Disk Utilization

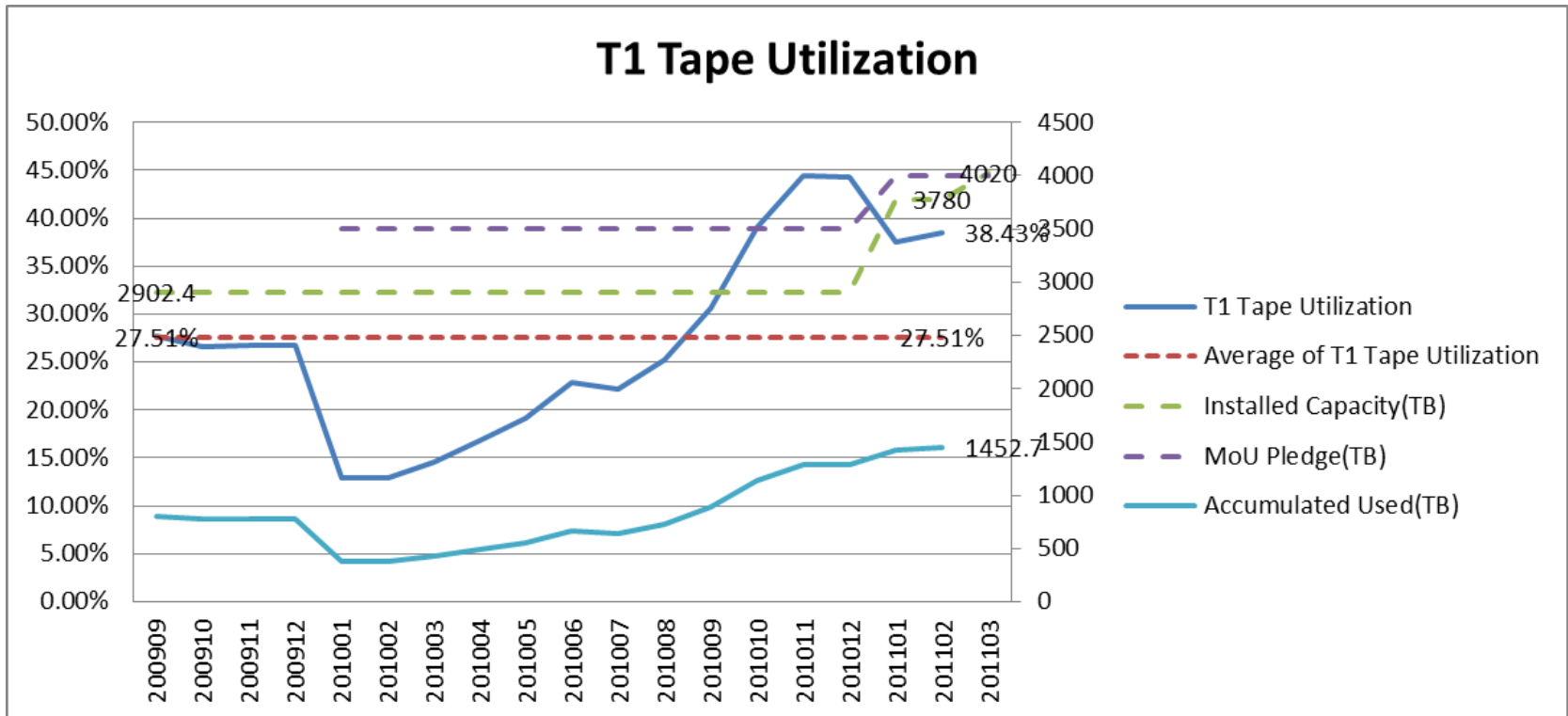


Average T1 Storage Utilization : **54.34%**

- Mar. 2011 – New procured 2PB disk online



T1 Tape Utilization



Average of T1 Tape Utilization : **27.51%**

- May 2010 – Tape HA with another robot were implemented
- Feb. 2011 – Reached 4PB tape capacity



ATLAS Transfers at ASGC

From 2010-01-01 to 2010-12-31:

1.73 PB flow into ASGC T1

605 TB flow into ASGC T2

Activity Summary ('2010-01-01 11:50' to '2010-12-31 15:50' UTC)

Click on the cloud name to view list of sites

| Cloud | Transfers | | | Registrations | | Errors | | | Services |
|----------------------|------------|------------|-----------|---------------|----------|----------|--------------|----------|----------|
| | Efficiency | Throughput | Successes | Datasets | Files | Transfer | Registration | Services | Grid |
| CA | 95% | 99 MB/s | 9218240 | 542566 | 9219454 | 506401 | 7 | 0 | |
| CERN | 85% | 186 MB/s | 9881503 | 317589 | 9869726 | 1714166 | 4 | 0 | |
| DE | 90% | 217 MB/s | 16685351 | 909112 | 16690679 | 1897459 | 13518 | 0 | |
| ES | 88% | 112 MB/s | 10131345 | 606085 | 10130853 | 1402192 | 0 | 0 | |
| FR | 90% | 220 MB/s | 21791938 | 1037095 | 21834943 | 2289979 | 16 | 0 | |
| IT | 82% | 148 MB/s | 10625979 | 597257 | 10610227 | 2297349 | 8641 | 0 | |
| ND | 80% | 77 MB/s | 6961894 | 296170 | 6958749 | 1729199 | 14650 | 0 | |
| NL | 82% | 151 MB/s | 7791421 | 710976 | 7785329 | 1708732 | 11455 | 0 | |
| TW | 92% | 74 MB/s | 5574676 | 334035 | 5560867 | 483644 | 0 | 0 | |
| UK | 87% | 152 MB/s | 17672301 | 1064947 | 17665461 | 2681984 | 0 | 0 | |
| US | 94% | 505 MB/s | 43216948 | 1319912 | 43389263 | 2823683 | 4401 | 0 | |



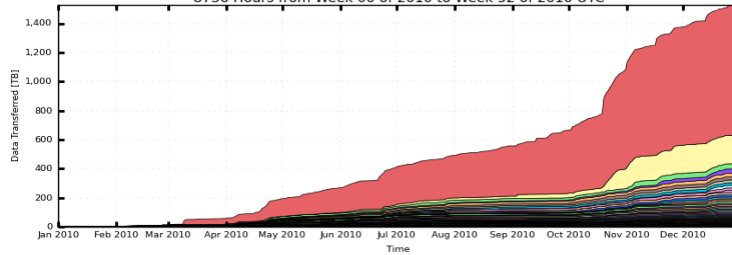
CMS Transfers at ASGC

From 2010-01-01 to 2010-12-31:

2.77 PB flow In/Out ASGC T1

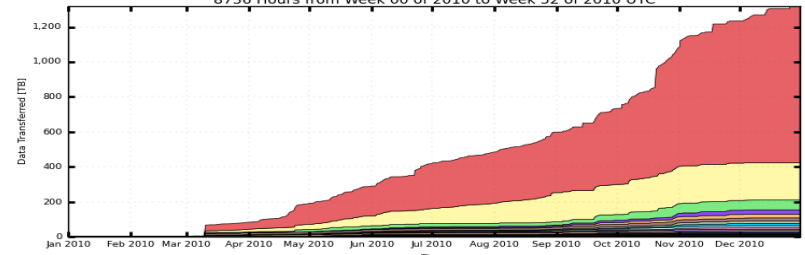
753 TB flow In/Out ASGC T2

CMS PhEDEx - Cumulative Transfer Volume
8736 Hours from Week 00 of 2010 to Week 52 of 2010 UTC



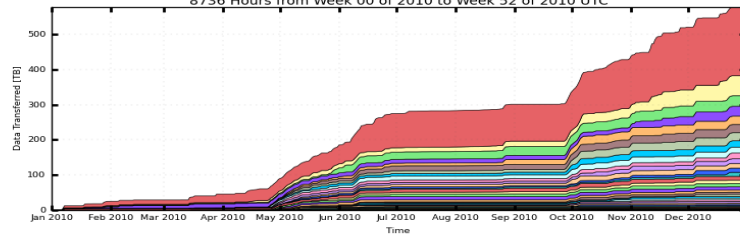
- | | | | | |
|---------------------|-------------------|-------------------|----------------------|-------------------|
| T1_TW_ASGC_MSS | T1_US_FNAL_Buffer | T3_US_FNALLPC | T2_US_Purdue | T2_US_MIT |
| T2_DE_DESY | T2_US_Nebraska | T2_US_Florida | T2_UK_London_IC | T2_US_Caltech |
| T2_RU_RRC_KI | T2_US_Wisconsin | T2_FR_GRIF_LLJ | T2_BE_UCL | T2_CH_CSCS |
| T2_ES_CIEMAT | T2_TR_METU | T2_ES_IFCA | T2_IT_Legnano | T2_TW_Taiwan |
| T2_IT_Pisa | T2_CH_CAF | T2_CN_Beijing | T2_US_UCSD | T1_IT_CNAF_Buffer |
| T2_FLHIP | T2_AT_Vienna | T2_FR_IPHC | T1_DE_KIT_Buffer | T2_DE_RWTH |
| T2_IT_Bari | T2_PT_NCG_Lisbon | T2_EE_Estonia | T2_KR_KNU | T2_PT_LIP_Lisbon |
| T2_BR_URJ | T2_PT_NCG_Lisbon | T2_RU_SINP | T1_FR_CCIN2P3_Buffer | T2_DE_WITH |
| T2_FR_GRIF_IRFU | T2_IN_TIFR | T2_UK_SGrid_RALPP | T1_UK_RAL_Buffer | T2_US_Florida |
| T2_UK_SGrid_Bristol | T2_FR_CCIN2P3 | T1_ES_PIC_Buffer | T2_PL_Warsaw | T2_US_Wisconsin |
- Total: 1,528 TB, Average Rate: 0.00 TB/s
... plus 14 more

CMS PhEDEx - Cumulative Transfer Volume
8736 Hours from Week 00 of 2010 to Week 52 of 2010 UTC



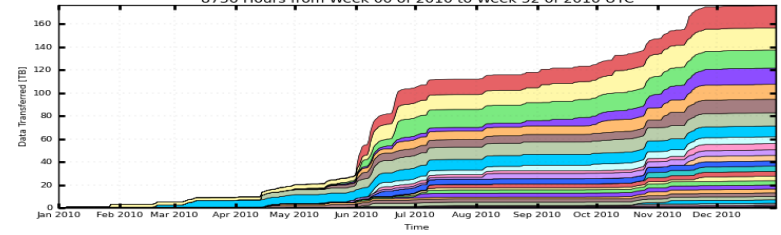
- | | | | | |
|-------------------|-------------------|-------------------|----------------------|---------------|
| T1_TW_ASGC_Buffer | T0_CH_CERN_Export | T1_US_FNAL_Buffer | T2_RU_JINR | T2_KR_KNU |
| T1_UK_RAL_Buffer | T2_TW_Taiwan | T2_IN_TIFR | T1_FR_CCIN2P3_Buffer | T2_RU_SINP |
| T1_DE_KIT_Buffer | T2_US_Caltech | T2_RU_JINR | T1_IT_CNAF_Buffer | T2_ES_CIEMAT |
| T2_ES_IFCA | T1_ES_PIC_Buffer | T1_CH_CERN_Buffer | T1_IT_CNAF_Buffer | T2_US_UCSD |
| T2_DE_RWTH | T2_RU_RRC_KI | T2_US_IHEP | T2_PT_NCG_Lisbon | T2_EE_Estonia |
| T2_US_Nebraska | T2_US_MIT | T2_US_Florida | T2_DE_DESY | T2_US_Purdue |
| T2_IT_Pisa | T2_IT_Bari | T2_US_Wisconsin | T2_US_Purdue | T2_CN_Beijing |
- Total: 1,318 TB, Average Rate: 0.00 TB/s

CMS PhEDEx - Cumulative Transfer Volume
8736 Hours from Week 00 of 2010 to Week 52 of 2010 UTC



- | | | | | |
|-------------------|-------------------|-------------------|----------------------|------------------|
| T1_US_FNAL_Buffer | T1_DE_KIT_Buffer | T1_IT_CNAF_Buffer | T1_FR_CCIN2P3_Buffer | T1_UK_RAL_Buffer |
| T2_DE_RWTH | T2_US_Nebraska | T2_US_UCSD | T2_FR_GRIF_LLJ | T2_US_Wisconsin |
| T3_US_FNALLPC | T2_UK_London_IC | T2_US_MIT | T2_FR_GRIF_IRFU | T1_ES_PIC_Buffer |
| T2_US_Caltech | T1_TW_ASGC_Buffer | T2_US_Purdue | T2_CH_CSCS | T2_DE_DESY |
| T2_CN_Beijing | T2_UK_SGrid_RALPP | T2_IT_Bari | T2_BE_IJHE | T2_US_Florida |
| T2_EE_Estonia | T2_ES_IFCA | T2_HU_Budapest | T2_FR_CCIN2P3 | T2_TR_METU |
| T2_US_Florida | T2_KR_KNU | T2_FR_IPHC | T2_TW_NCU | T2_BR_SPRACE |
| T2_PT_NCG_Lisbon | T2_BE_UCL | T2_AT_Vienna | T2_RU_JINR | T2_BR_SPRACE |
| T2_IT_Legnano | T2_BR_URJ | T2_UA_KIPT | T2_PT_NCG_Lisbon | T2_RU_RRC_KI |
| T1_CH_CERN_Buffer | T2_BR_URJ | | | |
| T2_RU_JINR | | | | |
- Total: 577.21 TB, Average Rate: 0.00 TB/s

CMS PhEDEx - Cumulative Transfer Volume
8736 Hours from Week 00 of 2010 to Week 52 of 2010 UTC



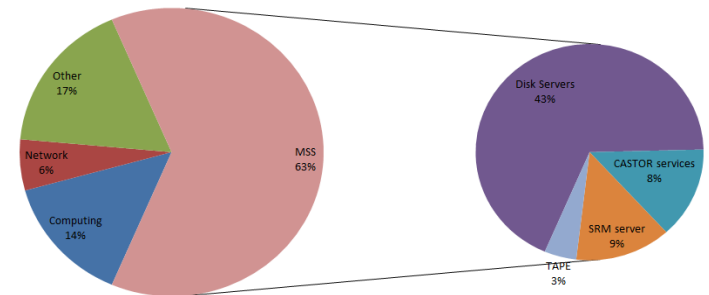
- | | | | | |
|-------------------|-------------------|---------------------|----------------------|----------------------|
| T2_US_Wisconsin | T3_US_FNALLPC | T2_US_Florida | T1_TW_ASGC_Buffer | T2_US_Nebraska |
| T2_US_Caltech | T2_DE_RWTH | T2_IT_Pisa | T2_FR_CCIN2P3_Buffer | T2_CH_CSCS |
| T2_CN_Beijing | T2_US_Wisconsin | T2_DE_IJHE | T2_RU_JINR | T2_DE_DESY |
| T2_US_MIT | T1_US_FNAL_Buffer | T2_UK_London_Brunel | T2_US_UCSD | T3_TW_NCU |
| T2_US_Purdue | T2_US_Purdue | T2_BR_URJ | T2_ES_IFCA | T2_ES_CIEMAT |
| T2_BR_SPRACE | T2_EE_Estonia | T2_UK_London_IC | T2_PT_NCG_Lisbon | T2_AT_Vienna |
| T2_UK_SGrid_RALPP | T2_FR_IPHC | T2_KR_KNU | T2_CH_CAF | T2_IT_Rome |
| T2_RU_RRC_KI | T2_IT_Bari | T2_US_Wisconsin | T2_FR_CCIN2P3 | T1_FR_CCIN2P3_Buffer |
- Total: 175.81 TB, Average Rate: 0.00 TB/s

Storage issue in 2010

- Over **50%** operation issues come from **storage system**.
- Disk servers configuration:
 - Bottleneck on those disk server with fewer CPU cores, memory and limited bandwidth.
 - Using one blade server connect to an array, the cost are high on blade server, array controller, rack space and power consumption.

ATLAS GGUS Tickets at ASGC T1

Type of GGUS Tickets (Jul - Oct 14, 2010)



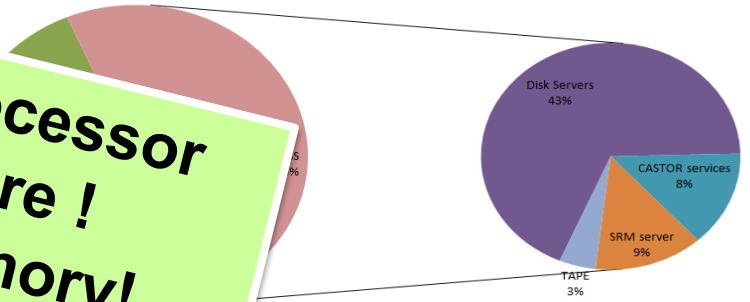
Storage issue in 2010

ATLAS GGUS Tickets at ASGC T1

Type of GGUS Tickets (Jul - Oct 14, 2010)

- Over 50% operation issues come from storage system.
- Disk servers configuration
 - Bottleneck on those server with fewer CPU, memory and limited bandwidth.
 - Using one blade server connect to an array, the cost are high on blade server, array controller, rack space and power consumption.

**Dual-processor
six-core !
24G Memory!
10Gb/s Ethernet!**





Storage issue in 2010

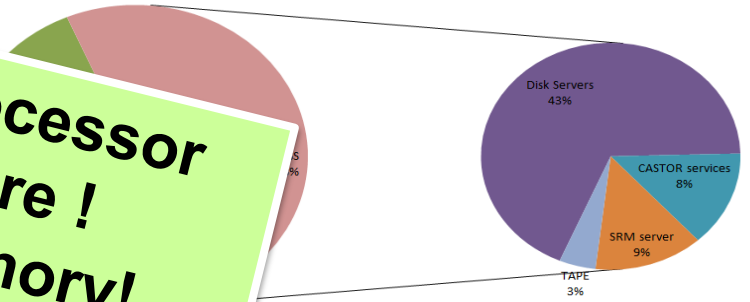
ATLAS GGUS Tickets at ASGC T1

Type of GGUS Tickets (Jul - Oct 14, 2010)

- Over 50% operation issues come from storage system.
- Disk servers configuration
 - Bottleneck on those server with fewer CPU, less memory and limited bandwidth.
 - Using one blade server connect to an array, cost are high on blade server, array controller, rack and power consumption.

**Dual-processor
six-core !
24G Memory!
10Gb/s Ethernet!**

**Dual Controller!
JBOD Expansion!**



Storage System Optimization

- Storage system upgrade:

- Disk server

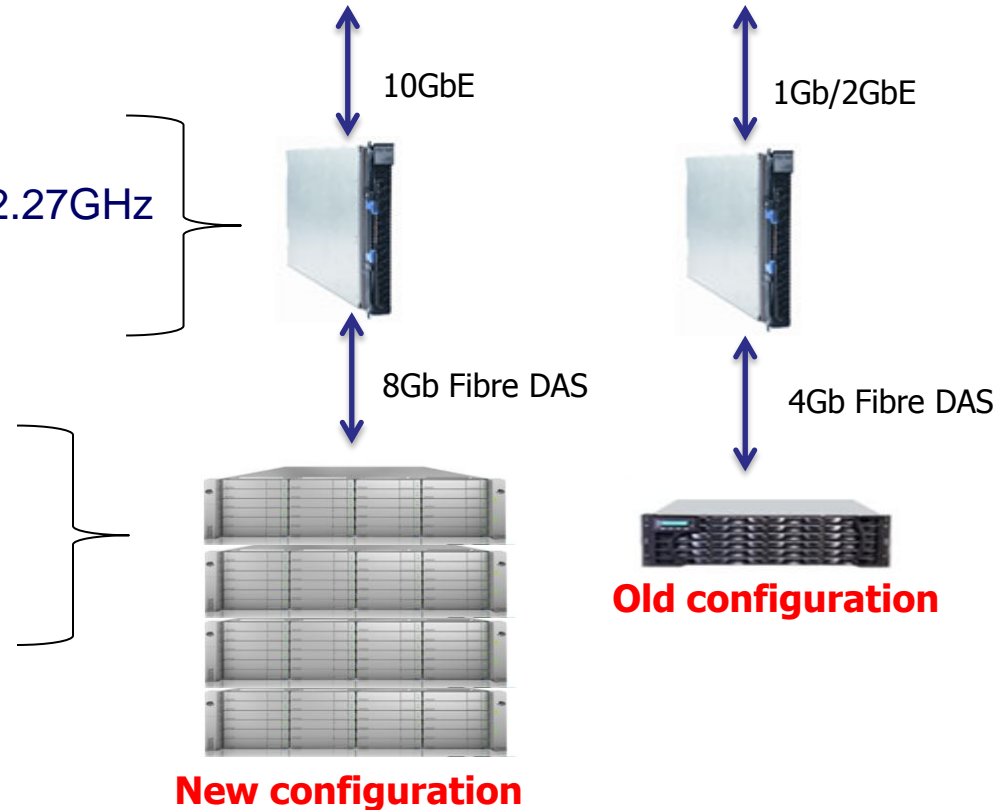
- IBM HS22 Blade
 - CPU: Intel Xeon CPU L5640 2.27GHz
 - Memory: 24G
 - 10GbE

- Backend arrays

- Dual controller
 - JBOD x 3, 6Gb SAS
 - 4U-24Bay x 4
 - 8Gb/s x 2 FC-Host

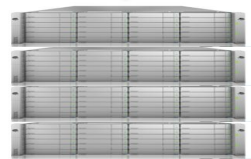
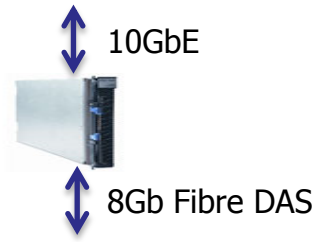
- Storage optimization:

- RAID5
 - Multipathing and LUN affinity
 - Tuned storage controller, OS (io-scheduler, TCP buffer...) and storage middleware (CASTOR and DPM)





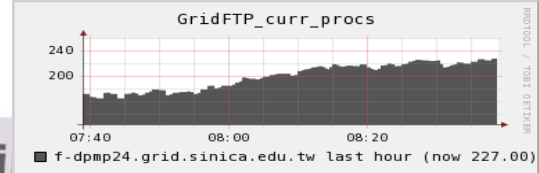
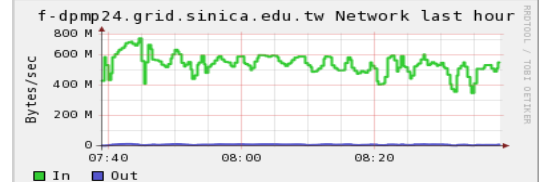
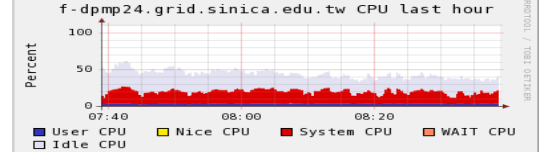
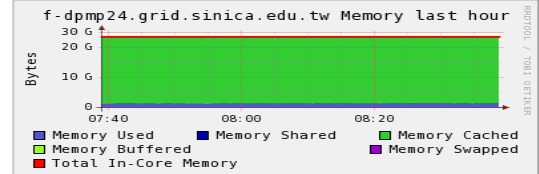
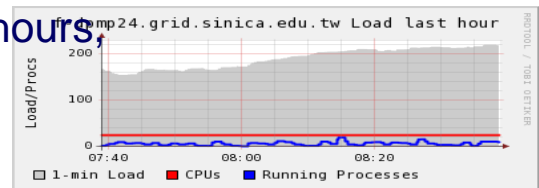
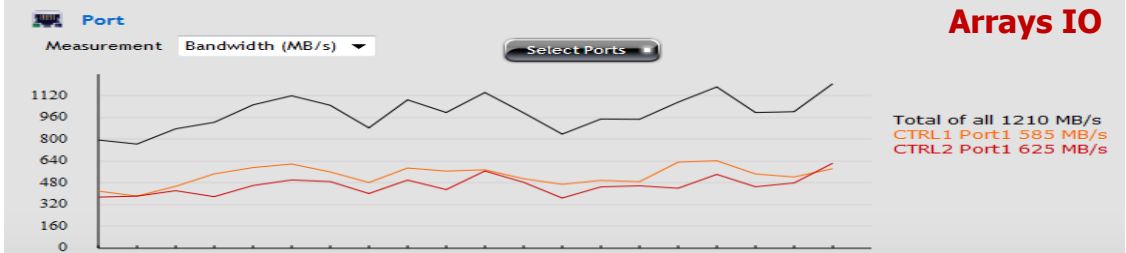
Disk server Performance



Observation while testing by ATLAS data transfers

- Peak network ~ **950MB/s**
- Peak arrays IO ~ **1210MB/s**
- Completed **4916** file transfers with **6.66TB** data flow in few hours,
2 errors (1 SRM_FAILURE; 1 globus_xio)
- It can handle more than **200** concurrent GridFTP transfers

| Time | IFACE | rxpck/s | txpck/s | rxbyt/s | txbyt/s | rxcmp/s | txcmp/s | rxmst/s |
|-------------|-------|-----------|----------|----------------------|------------|---------|---------|---------|
| 07:45:30 AM | eth2 | 629640.59 | 56313.86 | 952530045.54 | 3041432.67 | 0.00 | 0.00 | 2.97 |
| 07:45:31 AM | eth2 | 586384.00 | 46396.00 | 887168556.00 | 2512998.00 | 0.00 | 0.00 | 2.00 |
| 07:45:32 AM | eth2 | 572838.00 | 57892.00 | 866534085.00 | 3126657.00 | 0.00 | 0.00 | 3.00 |
| 07:45:33 AM | eth2 | 574588.89 | 50842.42 | 869258776.77 | 2756614.14 | 0.00 | 0.00 | 0.00 |
| 07:45:34 AM | eth2 | 666344.55 | 87286.14 | 1007847553.47 | 4713586.14 | 0.00 | 0.00 | 0.99 |
| 07:45:35 AM | eth2 | 545350.00 | 90933.00 | 824455144.00 | 4942066.00 | 0.00 | 0.00 | 0.00 |
| 07:45:36 AM | eth2 | 399546.00 | 93752.00 | 604046130.00 | 5077206.00 | 0.00 | 0.00 | 1.00 |





Storage Optimization Case I

Controller

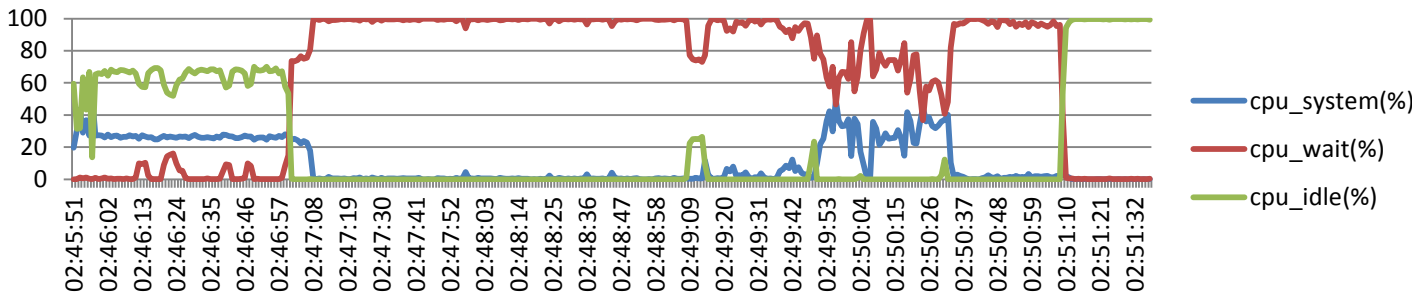
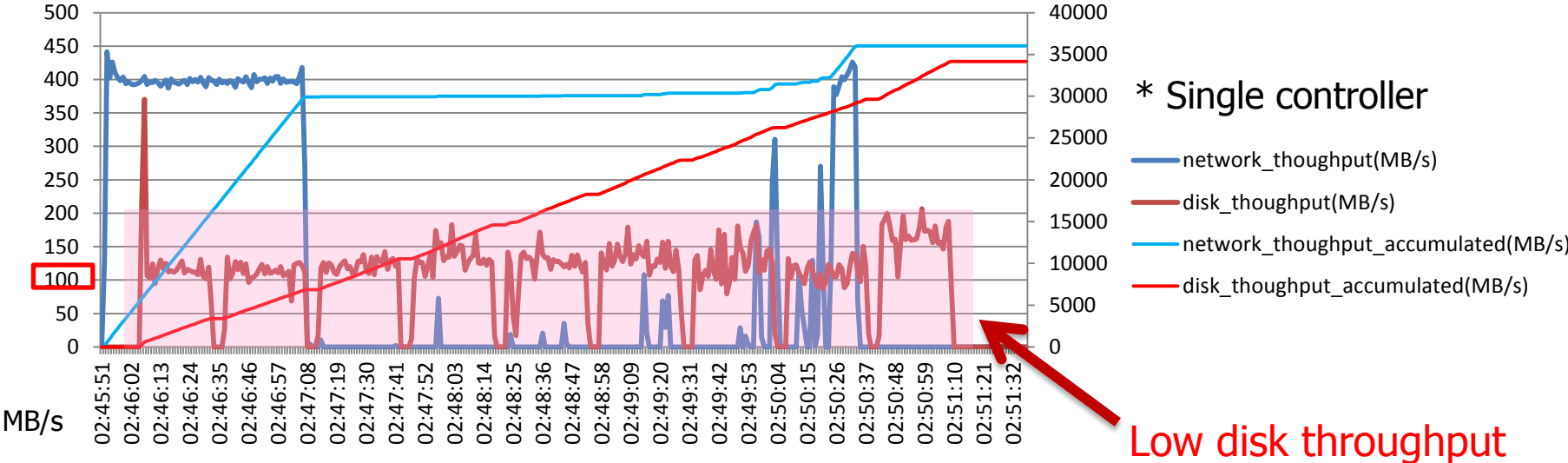


Write Test

- Each client copies a file **TO** the disk server
 - Using rfcop(rfio)
 - Each client copies a **DIFFERENT** file
 - **350** clients simultaneously start copying files

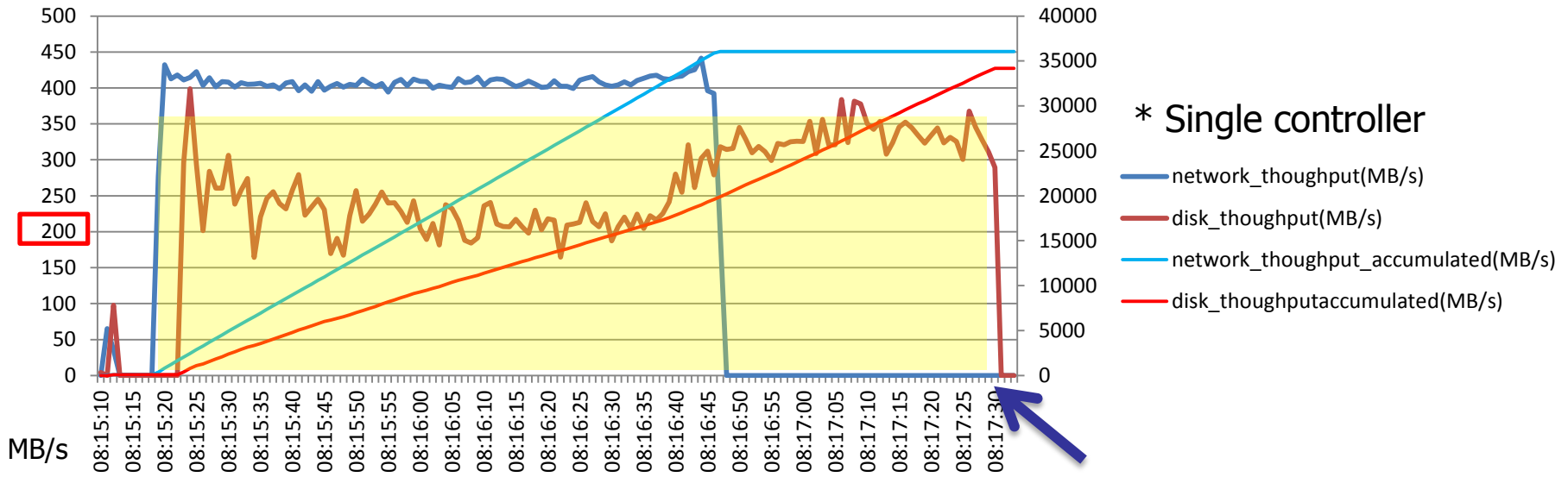


Low disk throughput – 100M x350

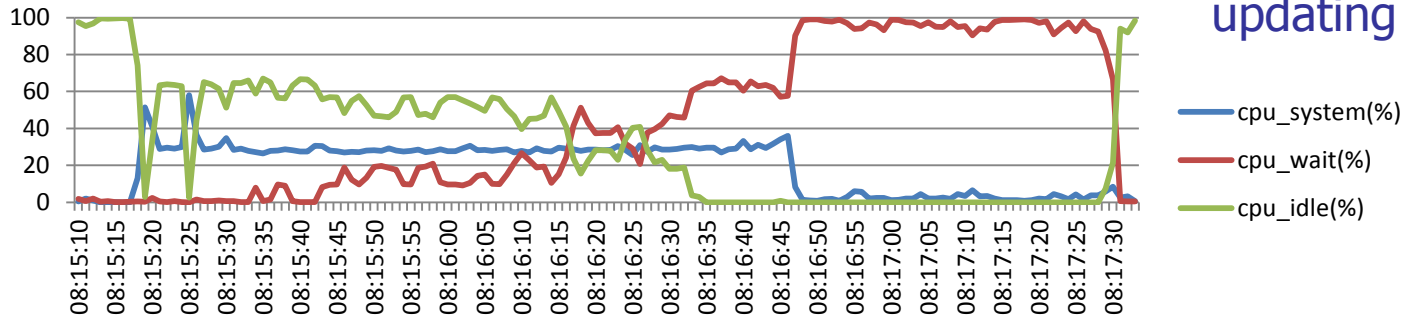




Improving disk throughput



Better disk throughput by updating controller firmware



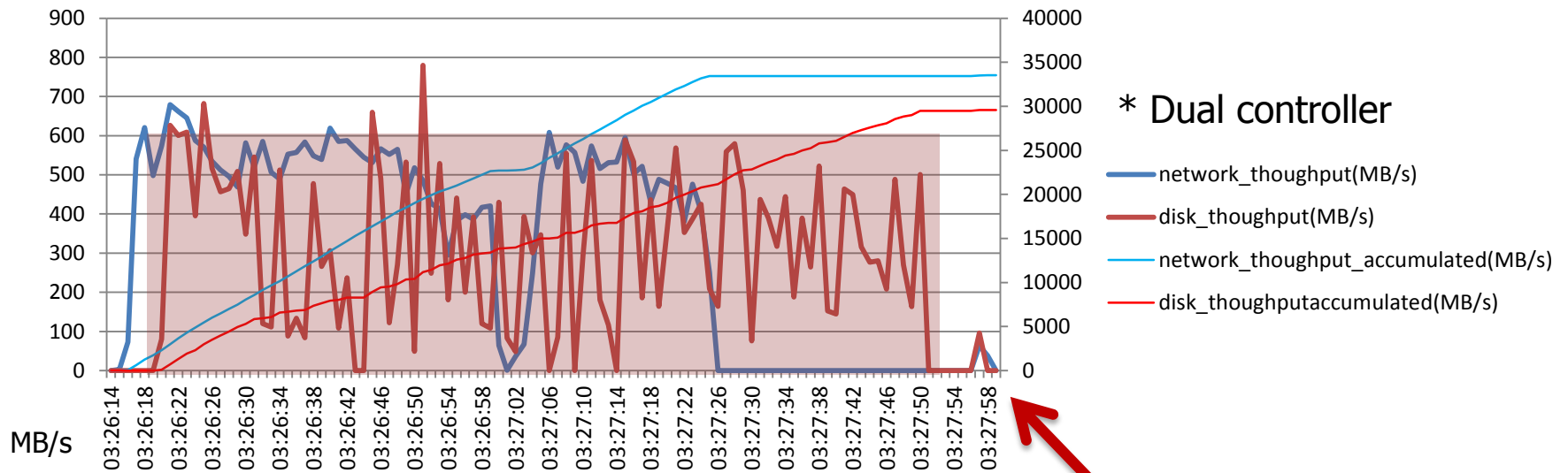


Storage Optimization Case II

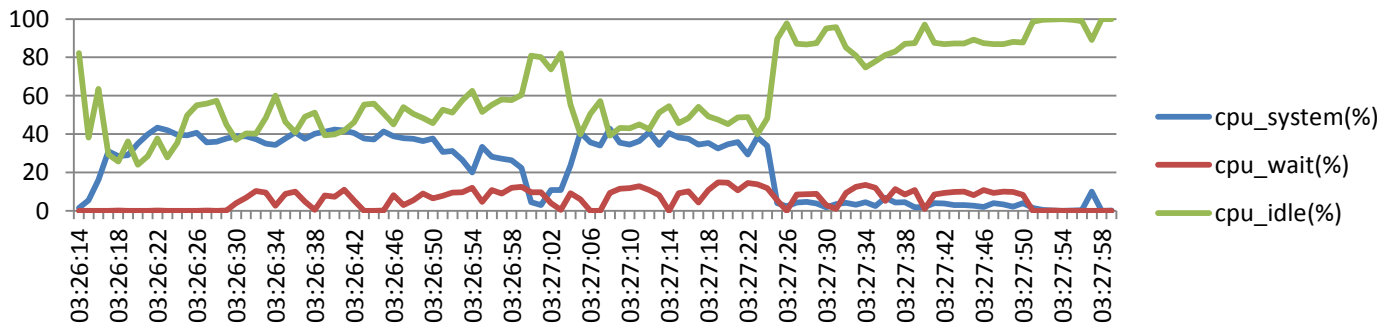
Multipath



Turbulent disk throughput –100M x350

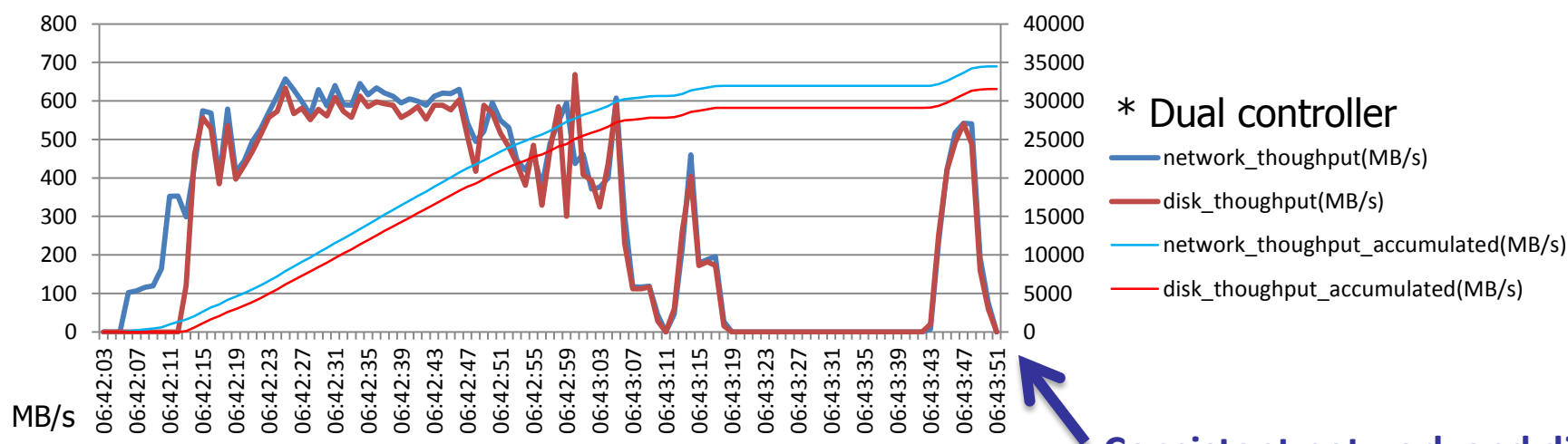


Turbulent disk throughput

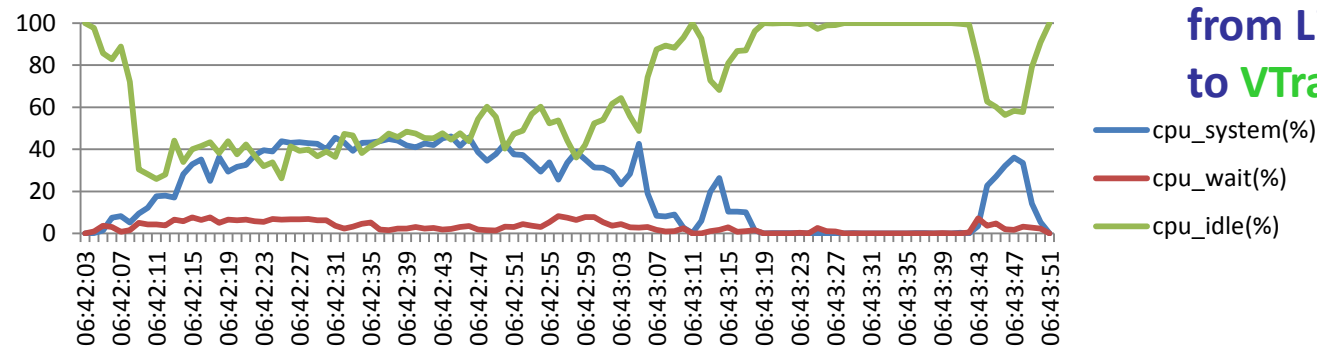




Improving Disk Throughput – 100M x350



Consistent network and disk throughput by changing from Linux default multipathing to VTrak multipathing





Multipath note

- Using VTrak multipathing and setting up ALUA
- Enable LUN affinity

```
devices {
  device {
    vendor          "Promise"
    product         "VTrak"
    path_grouping_policy group_by_prio
    getuid_callout  "/sbin/scsi_id -g -u -s /block/%n"
    prio_callout    "/sbin/mpath_prio_intel /dev/%n"
    path_checker    tur
    path_selector   "round-robin 0"
    hardware_handler "1 alua"
    failback        immediate
    rr_weight       uniform
    rr_min_io       100
    no_path_retry   queue
    features        "1 queue_if_no_path"
    product_blacklist "VTrak V-LUN"
  }
}
```

```
CLU | Controller 1 Settings
CtrlId          : 1
Alias           : 
LUN Affinity    : Enabled
```

```
CLU | Logical Drive 0 Info and Settings
LdId            : 0                ArrayId         : 0
OperationalStatus : OK
RAIDLevel       : RAID5           PreferredCtrlId : 1
NumOfUsedPD     : 12              NumOfAxles      : 1
Stripe         : 64 KB           Sector          : 512 Bytes
Capacity        : 20.01 TB        PhysicalCapacity : 21.83 TB
SYNCed          : Yes             CurrentWritePolicy: WriteBack
Alias           : LD0
ReadPolicy      : ReadAhead       WritePolicy      : WriteBack
```

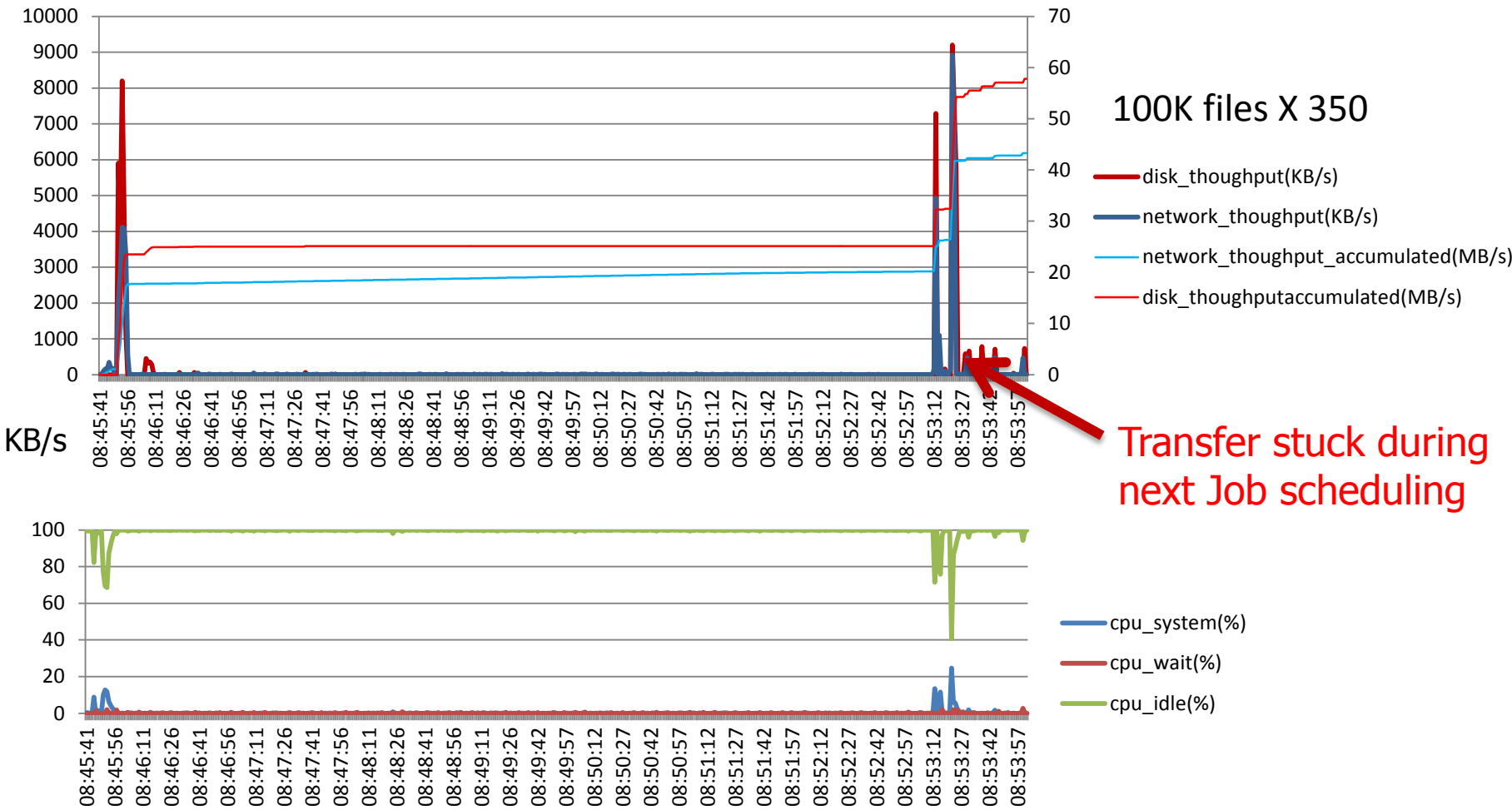


Storage Optimization Case III

CASTOR and DPM

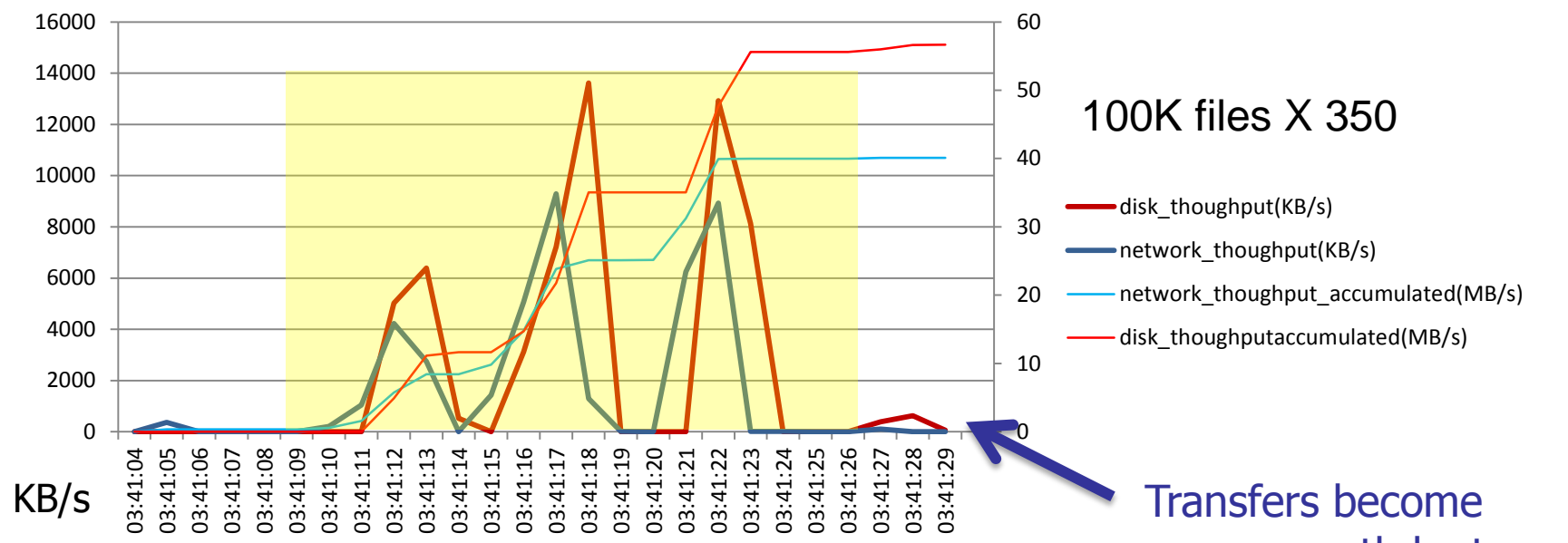


CASTOR- Poor efficiency in writing small size files

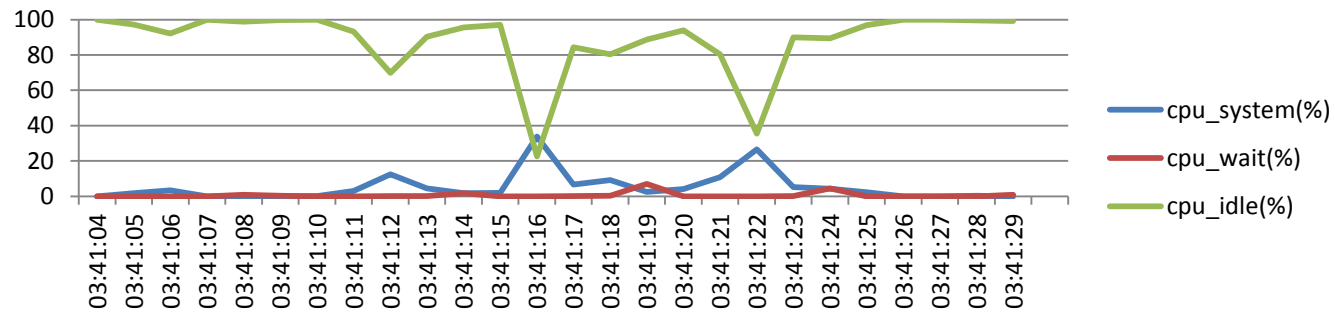




CASTOR LSF Tuned



Transfers become more smooth by tune CASTOR LSF



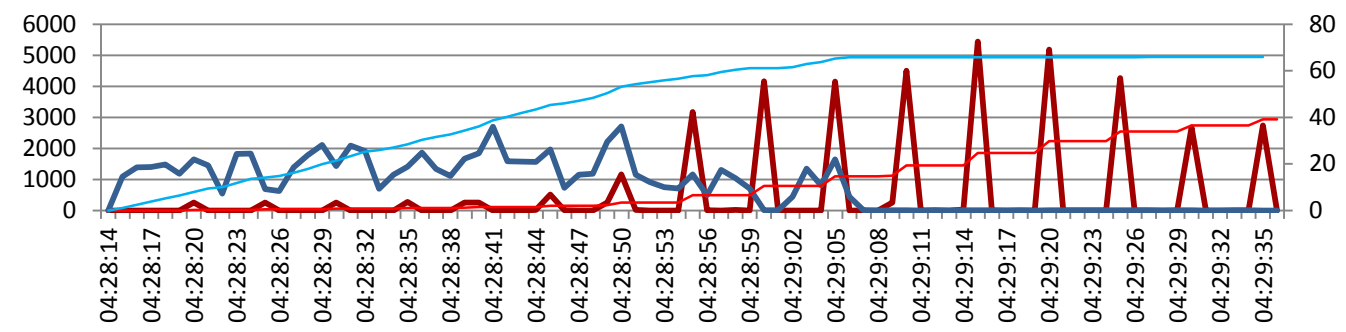


LSF tuning note

- `lsb.queues`
 - `NEW_JOB_SCHED_DELAY`
 - `CHUNK_JOB_SIZE`



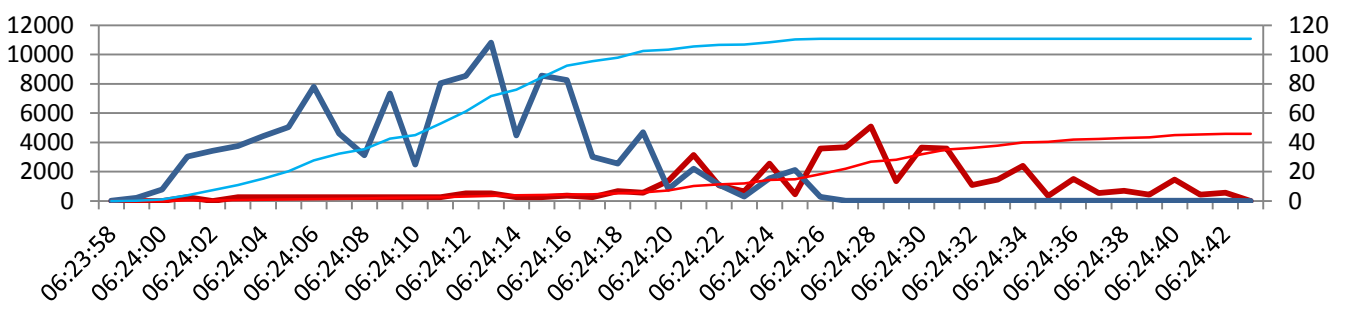
DPM Tuning- 100K x 350



DPNS threads 20

Duration: 1m 22s

- disk_throughput(KB/s)
- network_throughput(KB/s)
- disk_throughput_accumulated(MB/s)
- network_throughput_accumulated(MB/s)

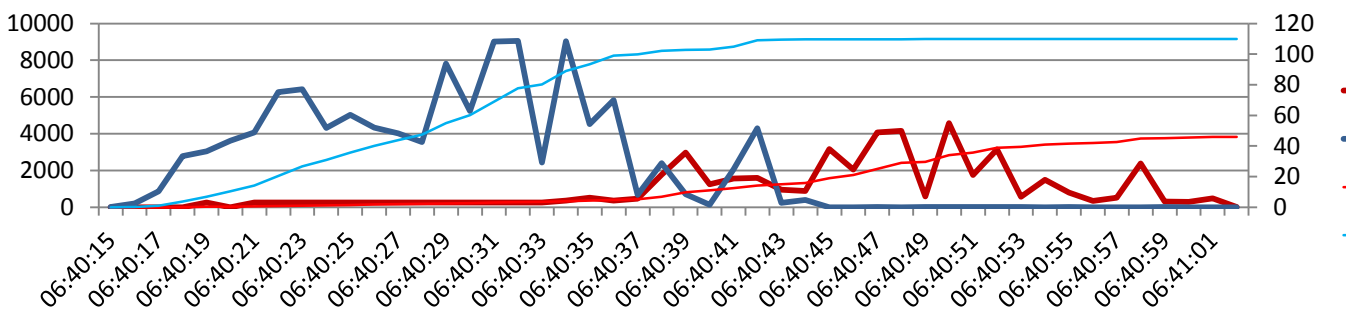


DPNS threads 50

Duration: 44s



- disk_throughput(KB/s)
- network_throughput(KB/s)
- disk_throughput_accumulated(MB/s)
- network_throughput_accumulated(MB/s)



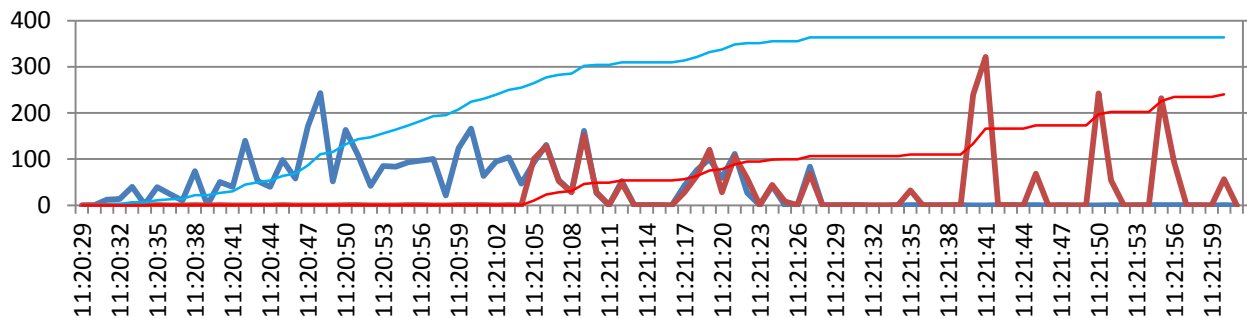
DPNS threads 80

Duration: 47s

- disk_throughput(KB/s)
- network_throughput(KB/s)
- disk_throughput_accumulated(MB/s)
- network_throughput_accumulated(MB/s)



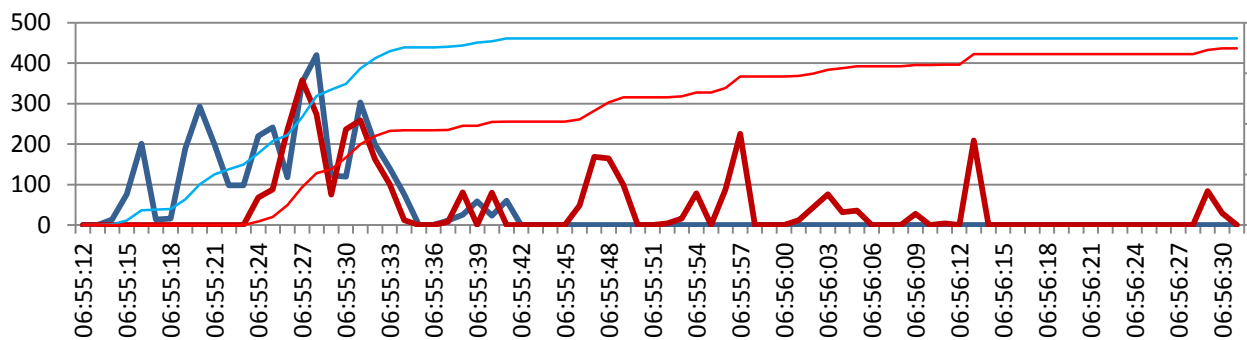
DPM Tuning- 1M x 350



DPNS threads 20

Duration: 1m 32s

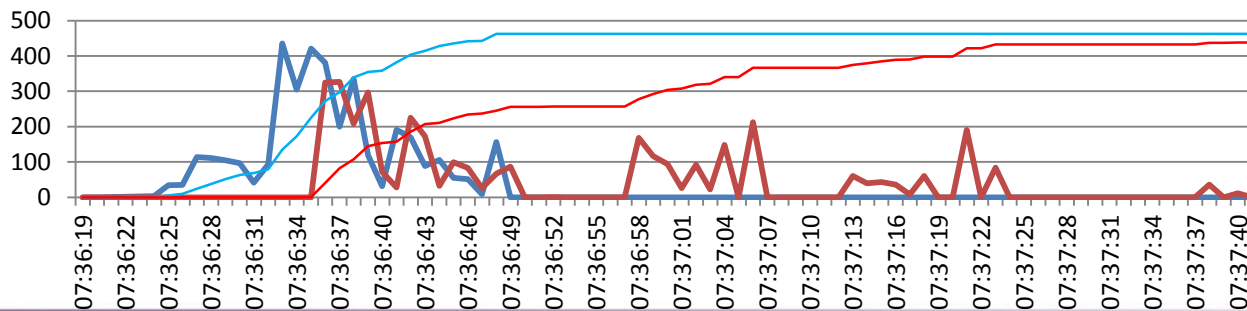
- network_thoughput(MB/s)
- disk_thoughput(MB/s)
- network_thoughput_accumulated(MB/s)
- disk_thoughputaccumulated(MB/s)



DPNS threads 50

Duration: 1m 18s

- network_thoughput(MB/s)
- disk_thoughput(MB/s)
- network_thoughput_accumulated(MB/s)
- disk_thoughputaccumulated(MB/s)



DPNS threads 80

Duration: 1m 18s

- network_thoughput(MB/s)
- disk_thoughput(MB/s)
- network_thoughput_accumulated(MB/s)
- disk_thoughputaccumulated(MB/s)



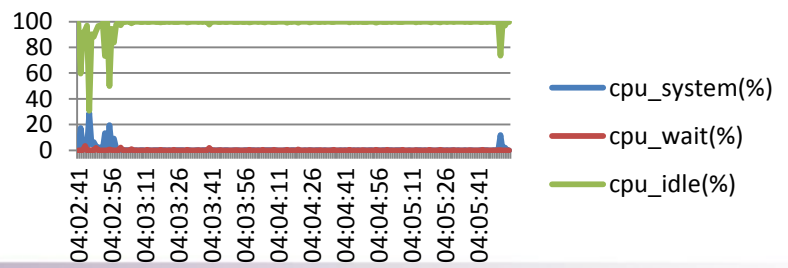
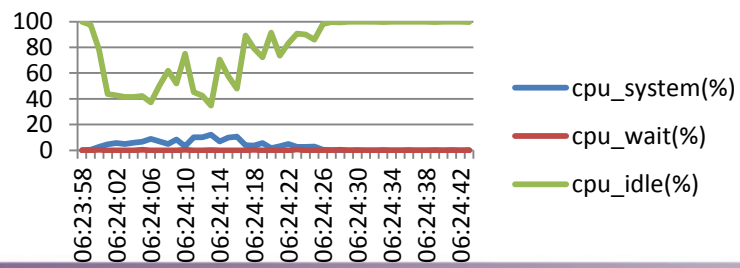
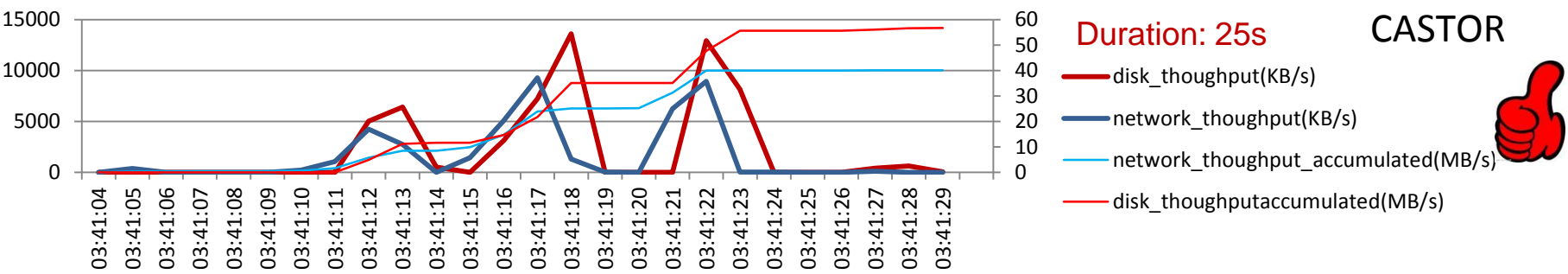
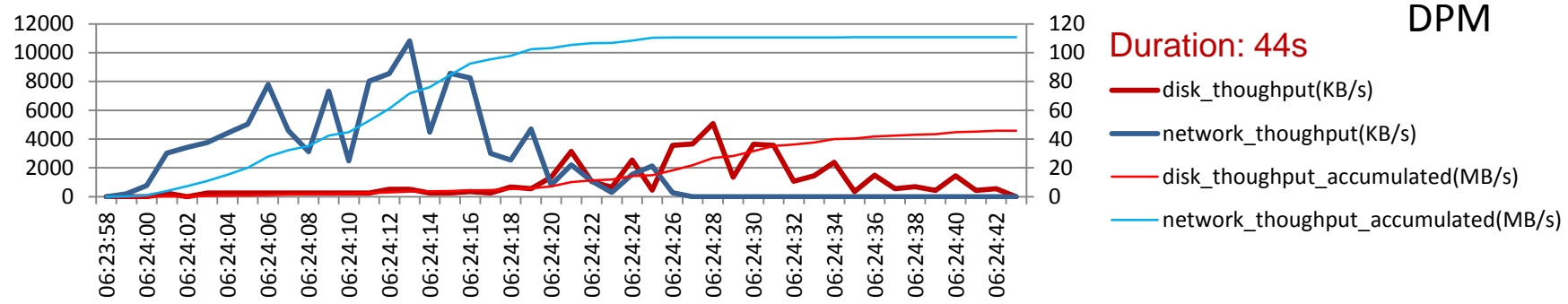
DPM tuning note

/etc/sysconfig/dpnsdaemon

```
[root@t-dpm ~]# grep [a-z,A-Z] /etc/sysconfig/dpnsdaemon
# $Id: dpnsdaemon.sysconfig.mysql,v 1.13 2007/07/26 12:09:10 slemaitr Exp $
# @(#) $RCSfile: dpnsdaemon.sysconfig.mysql,v $ $Revision: 1.13 $ $Date: 2007/07/26 12:09:10 $ CERN/IT/AD
C/CA Jean-Damien Durand
# should the DPNS daemon run?
# any string but "yes" will be equivalent to "no"
RUN_DPNSDAEMON="yes"
# should the DPNS be read-only ?
# any string but "yes" will be equivalent to "no"
RUN_READONLY="no"
# should we run with another limit on the number of file descriptors than the default?
# any string will be passed to ulimit -n
#ULIMIT_N=4096
# Change and uncomment the variables below if your setup is different than the one by default #
#ALLOW_COREDUMP="yes"
# DPNS variables #
# - Number of DPNS threads :
#NB_THREADS=20
# - DPNS log file :
#DPNSDAEMONLOGFILE="/var/log/dpns/log"
# - DPNS configuration file :
export DPNS_HOST=t-dpns.grid.sinica.edu.tw
export DPM_HOST=t-dpm.grid.sinica.edu.tw
NSCONFIGFILE=/opt/lcg/etc/NSCONFIG
ORACLE_HOME=
```

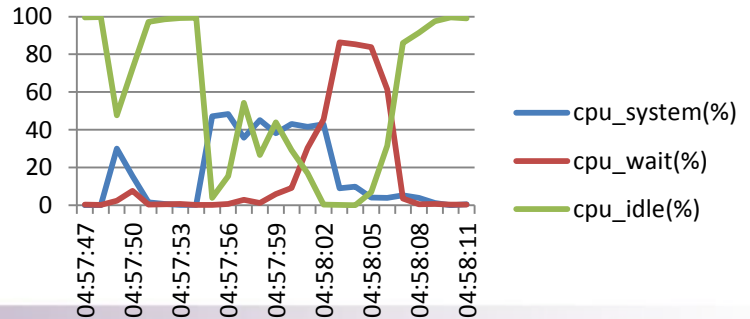
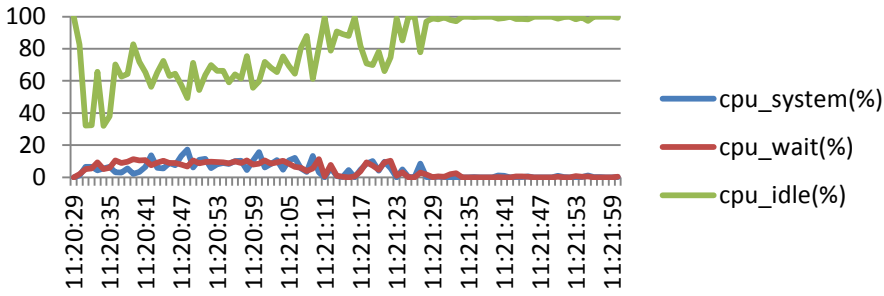
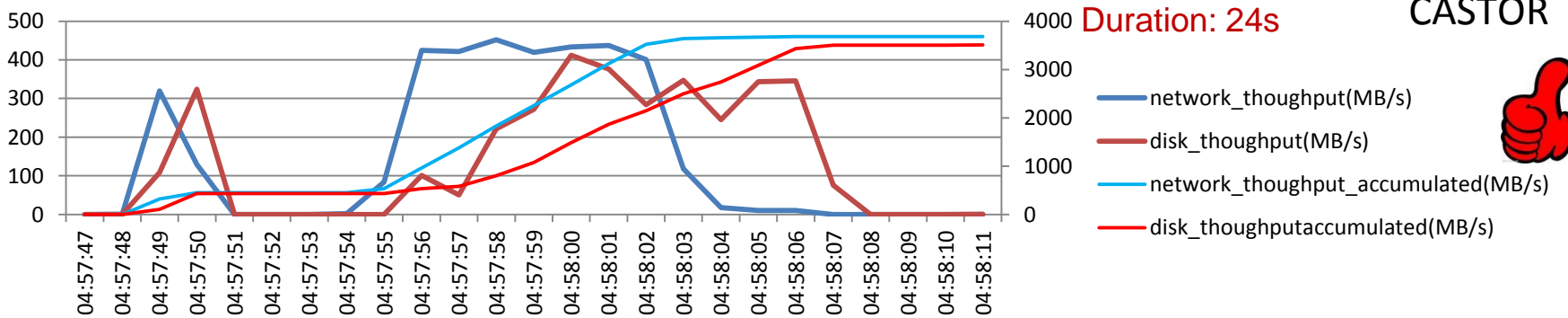
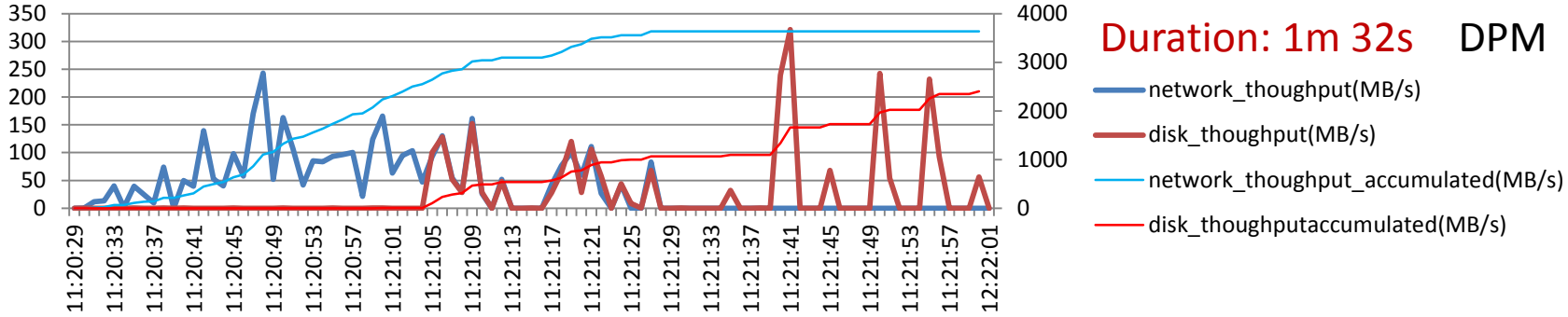


DPM V.S. CASTOR- 100K x 350



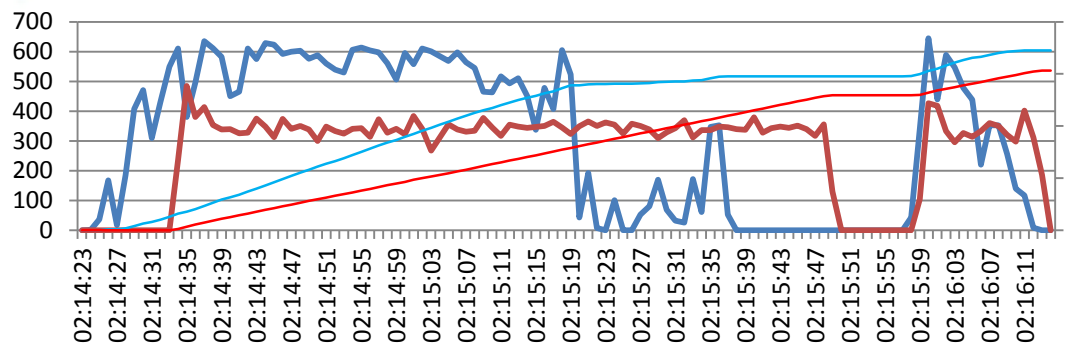


DPM V.S. CASTOR- 10M x 350



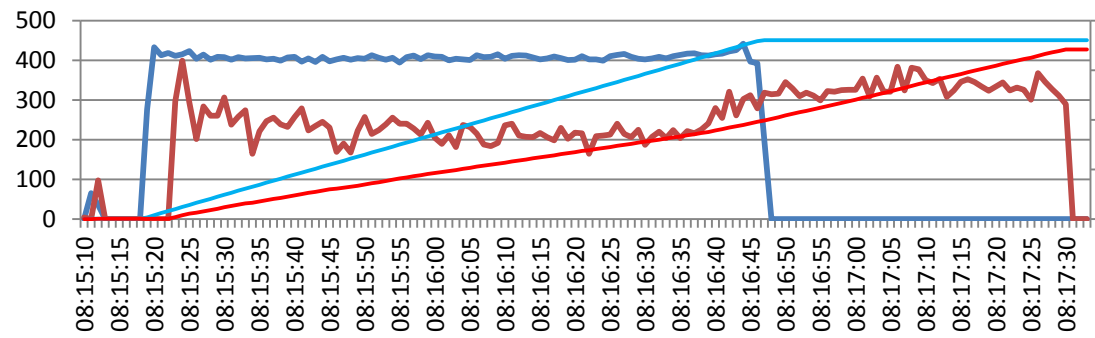


DPM V.S. CASTOR- 100M x 350



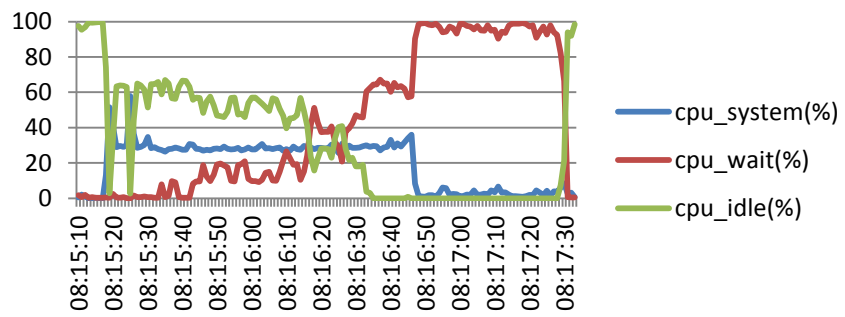
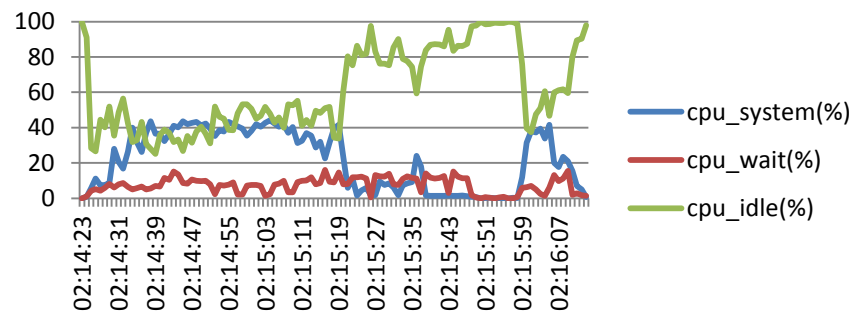
Duration: 1m 51s DPM

- network_throughput(MB/s)
- disk_throughput(MB/s)
- network_throughput_accumulated(MB/s)
- disk_throughput_accumulated(MB/s)



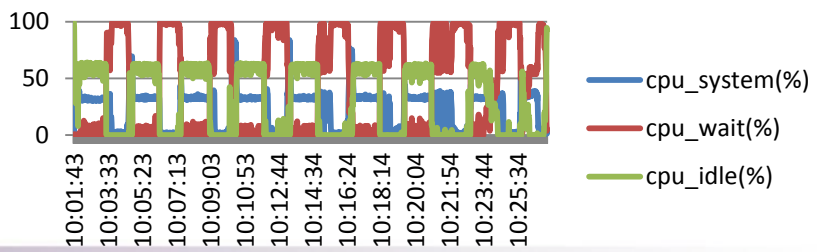
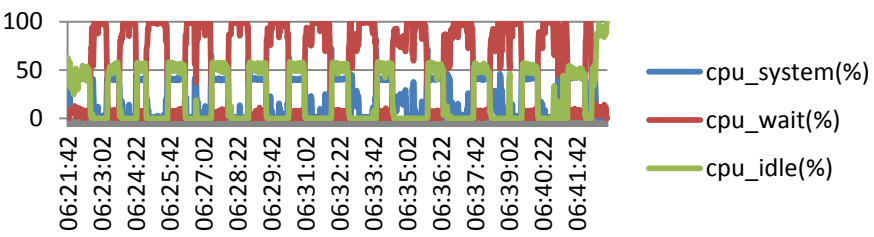
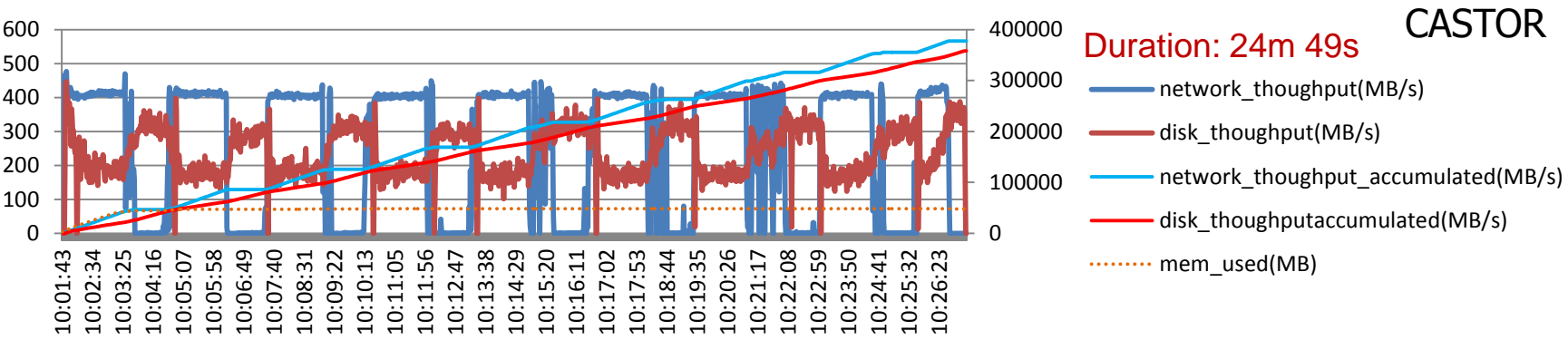
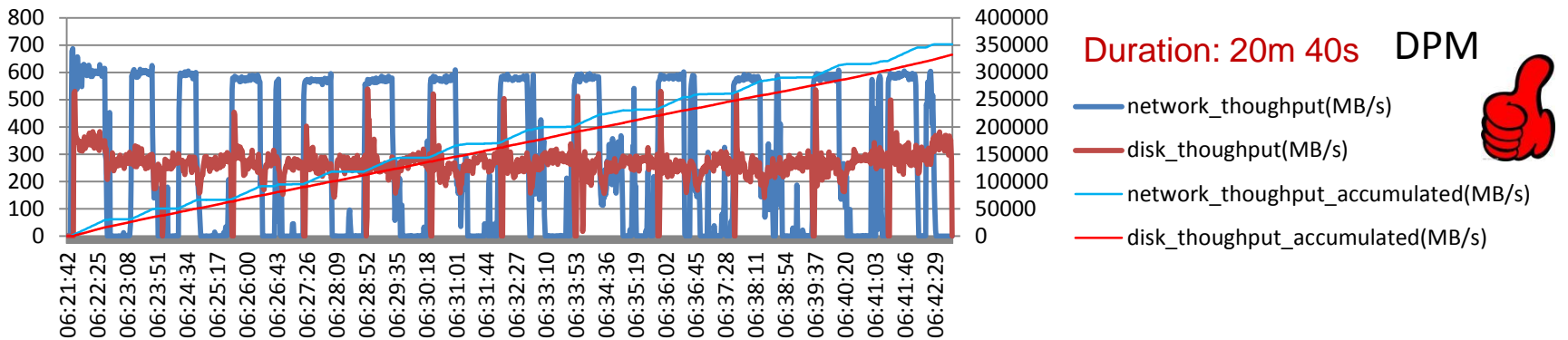
Duration: 2m 21s CASTOR

- network_throughput(MB/s)
- disk_throughput(MB/s)
- network_throughput_accumulated(MB/s)
- disk_throughputaccumulated(MB/s)





DPM V.S. CASTOR – 1G x 350





Summary and Next Step

- Optimized configuration based on recent hardware was able to reach close to full speed performance.
- Improvements in several aspects by new configuration:
 - Better throughput (~950MB/s in peak) has been seen during testing
 - Save Rack space and Power
 - 7U blade chassis can handle more than 2PB spaces
 - 21000W V.S 2900W
 - Save Money
 - Reduce #controllers and blade servers
- New configuration with 2.3PB were online production on Mar. 17, keep monitoring to ensure the performance.
- Next step --> reconfigure and optimize of old disk arrays



Thank You for Your Attention!