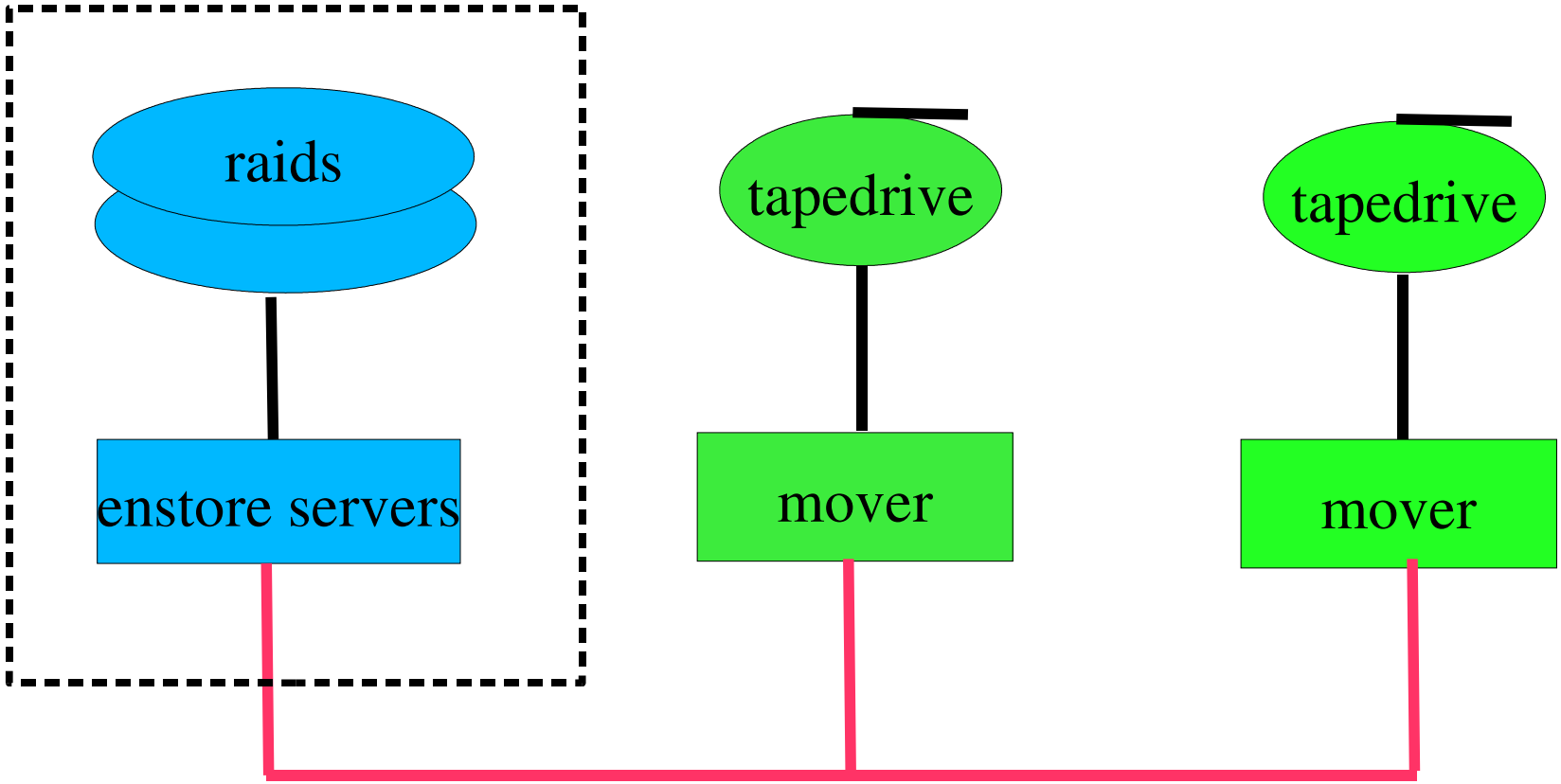
Strategically Deployable Enstore
(AKA ScrewDriverless Enstore)
Project Status Report
August 7, 2007

Outline

- objectives
- hardware configuration
- system installation
- software
- moving services
- milestones
- deployment plan
- staffing
- conclusion

Current Enstore HW configuration



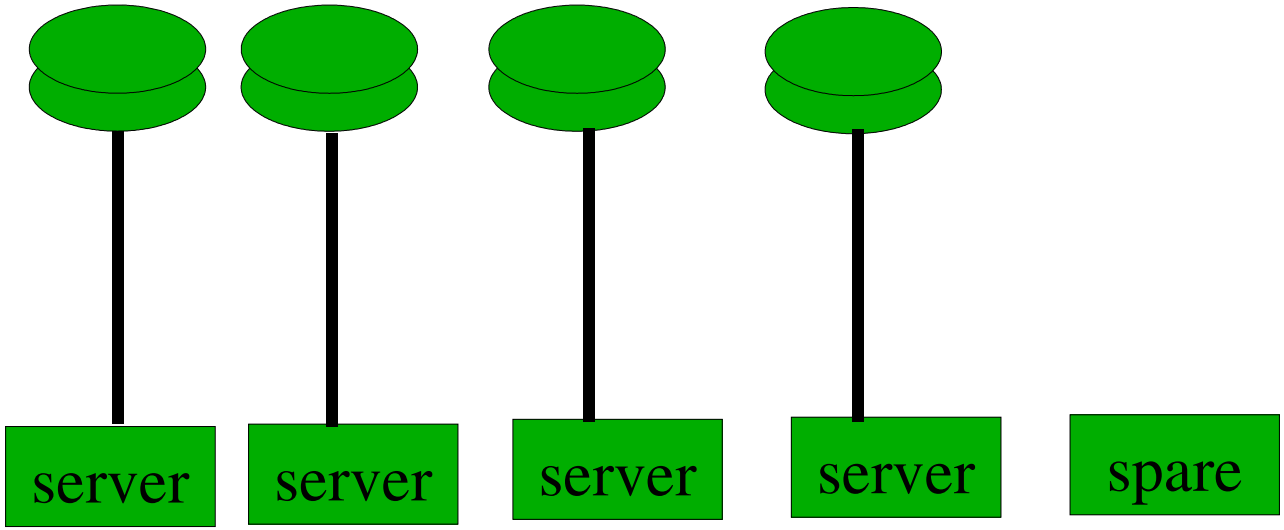
Objectives

- Change the support model so that hardware repairs/replacements can be made next day under warranty
- Reduce the cost and downtime associated with upgrading the systems
- Portable services able to run on any enstore server.

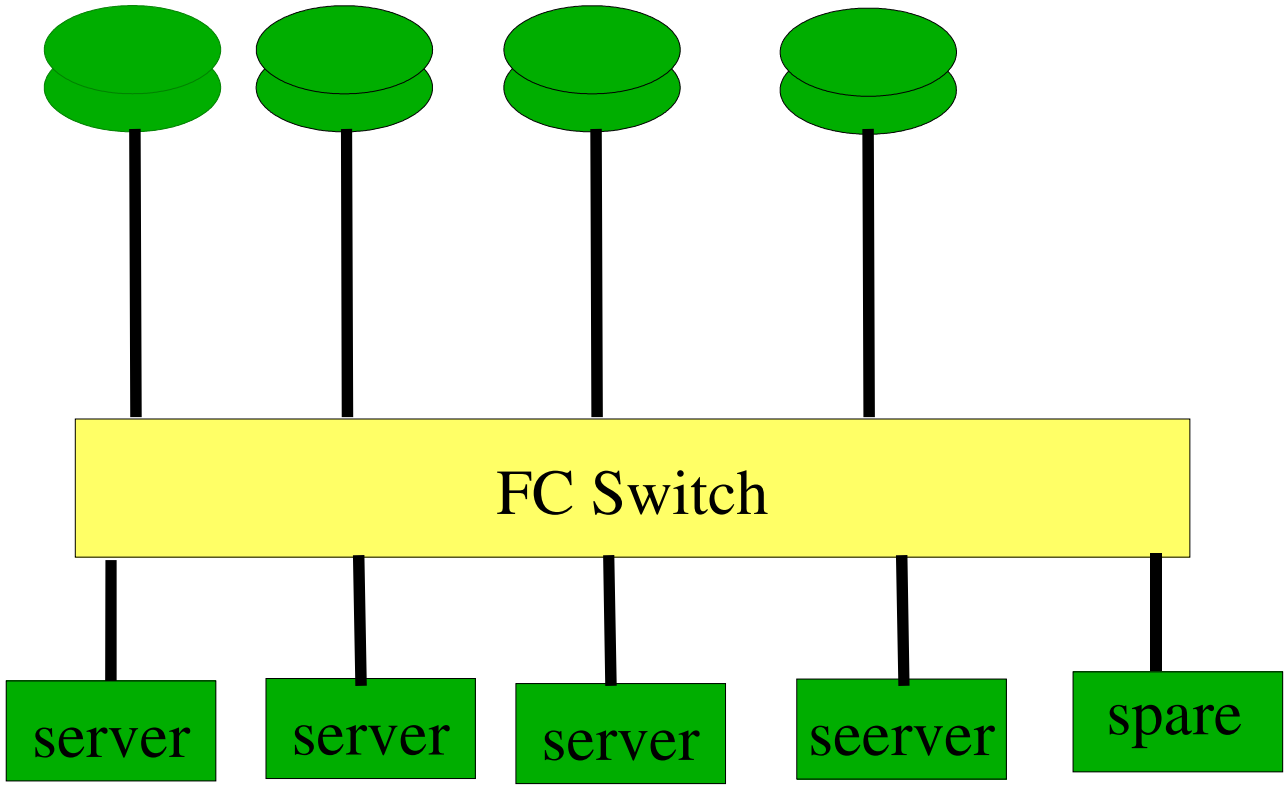
Selection of Hardware Configuration

- Current configuration: server with directly attached raid. Service is tied to server if it uses data on raid.
- For SDE we need hardware configuration that provides the capability to connect any enstore raid to any server.
- We considered two options (see pictures):
  - configuration with dedicated FC switch on which servers and raids are attached - Dedicated Storage.
  - Publicly available SAN (BlueArc) – Shared Storage

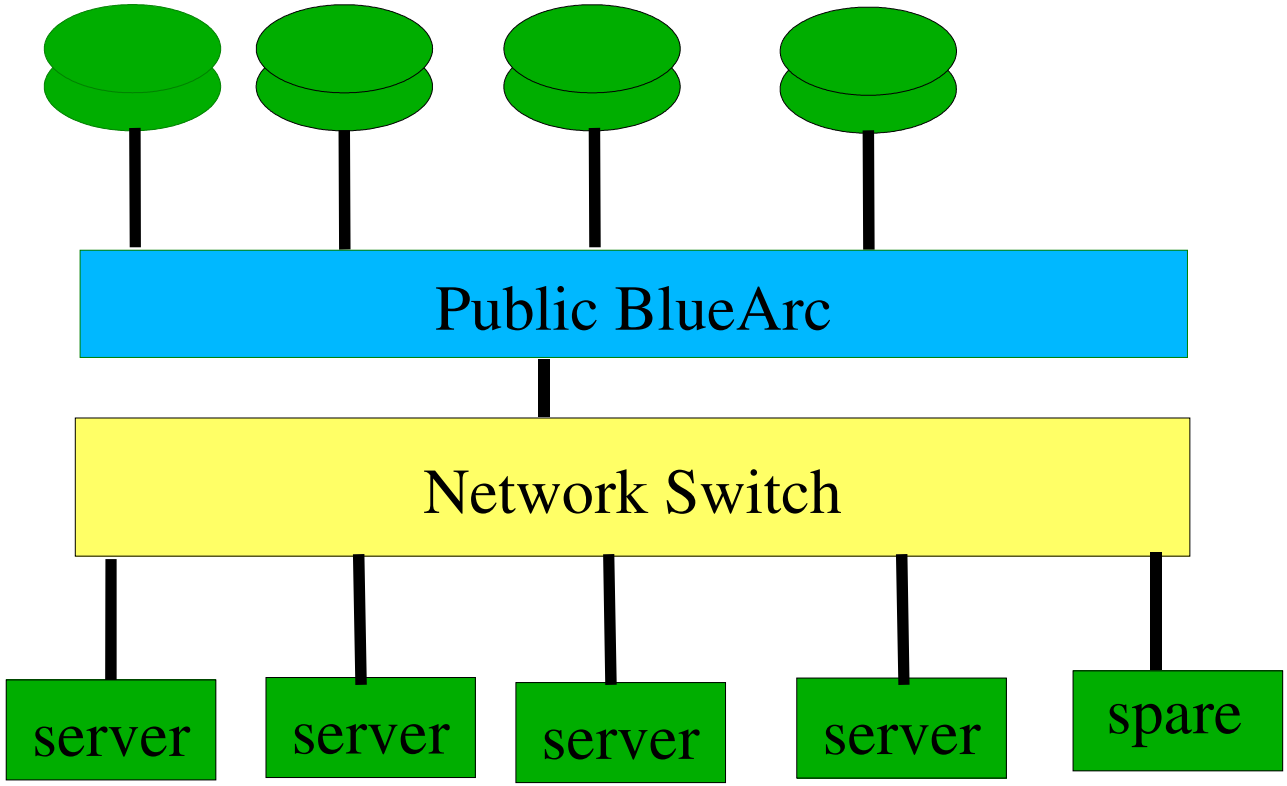
Current Hardware Configuration



FC Attached Raids



Raids via BlueArc SAN



Dedicated vs. Shared

Dedicated Storage

- pros.
  - predictable throughput
  - in general this configuration has a higher transfer rates limit
- cons.
  - administration load on SSA
  - no automatic redundancy

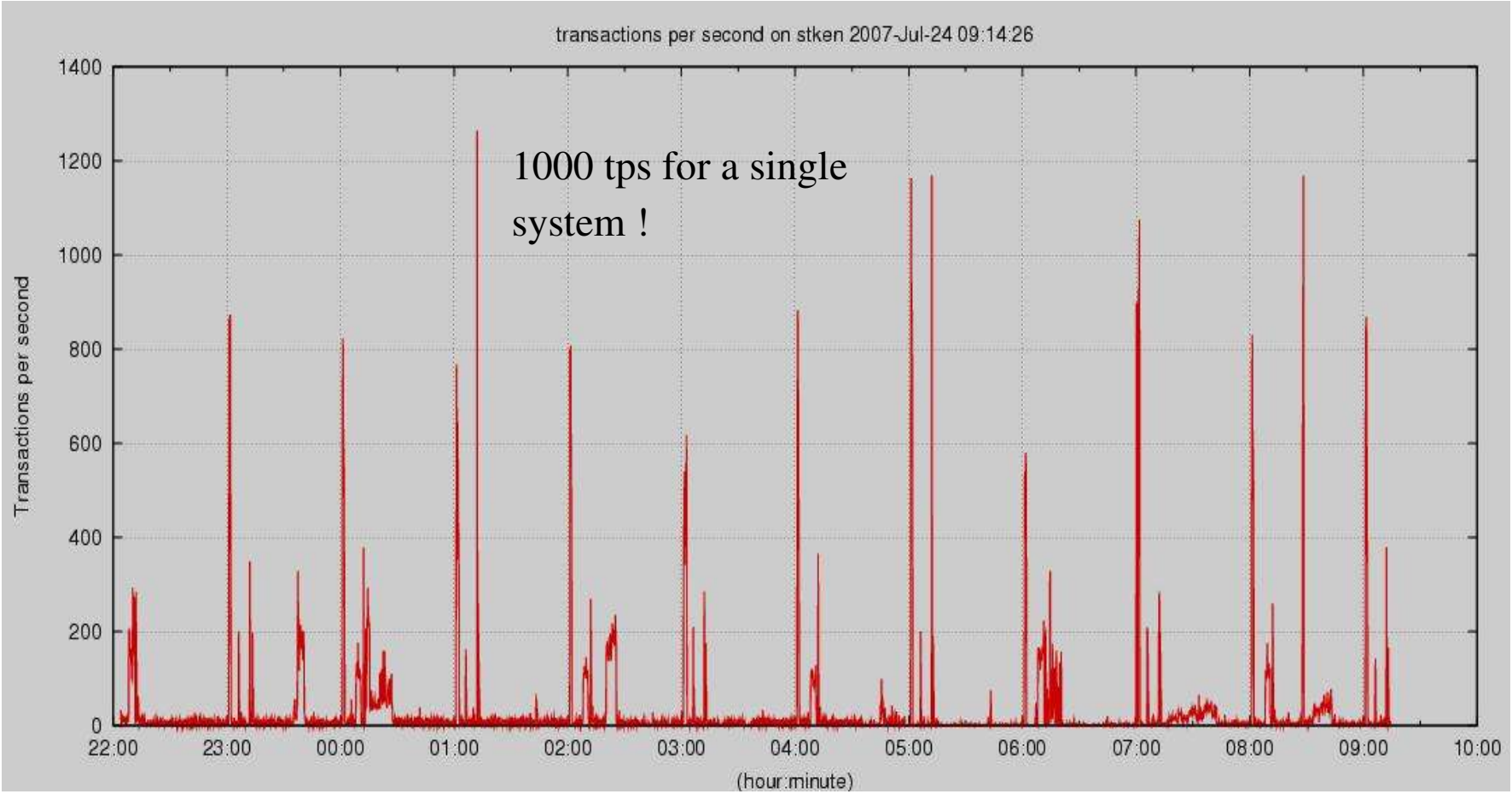
Shared Storage

- pros.
  - less administration load on SSA
  - automatic redundancy
- cons.
  - not predictable throughput
  - potential degradation or failure of service due to unpredictable system load

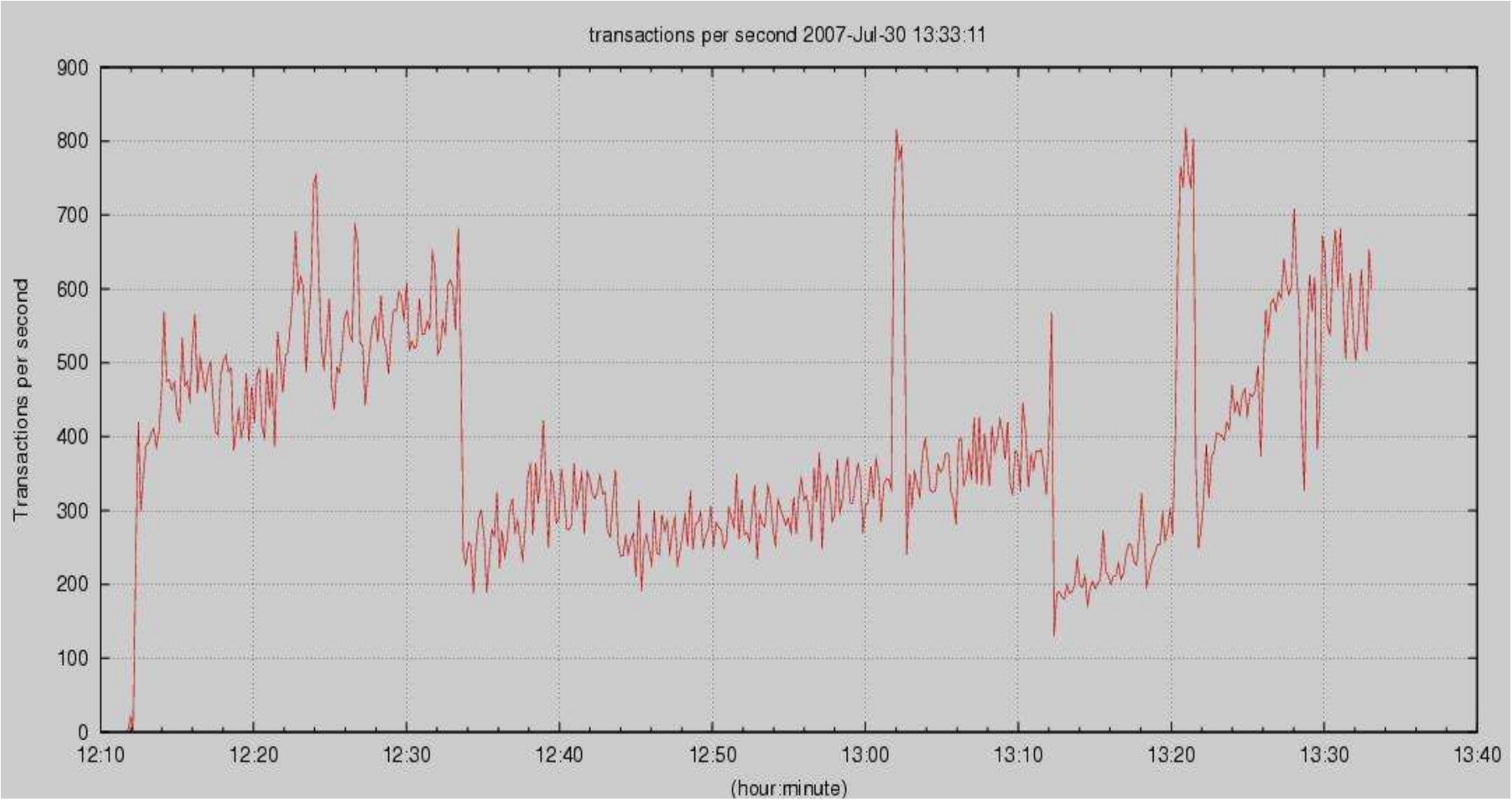
Evaluation of Shared Storage

- In our evaluation of BlueArc we could not achieve the rates we have on the production systems with the level of support we had. We expect to have a certain bandwidth available all the time. Public system does not show performance adequate to one for current production systems (see plots below)
- We decided to go with a private SAN
- We tested redundant switch configuration.
- A separate node on a public SAN will be used for SW installation and configuration of enstore nodes.

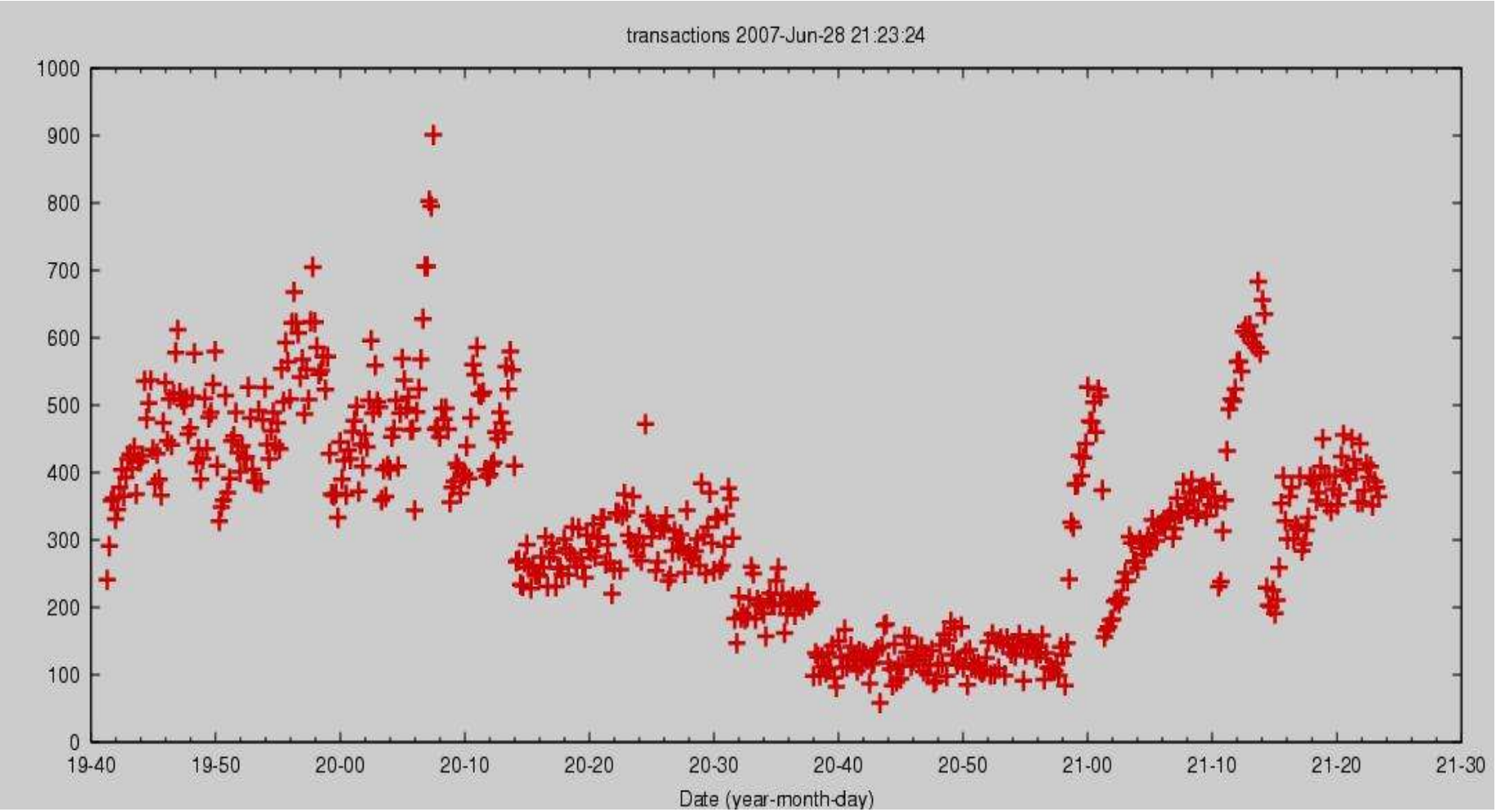
pnfs transaction rates for production system



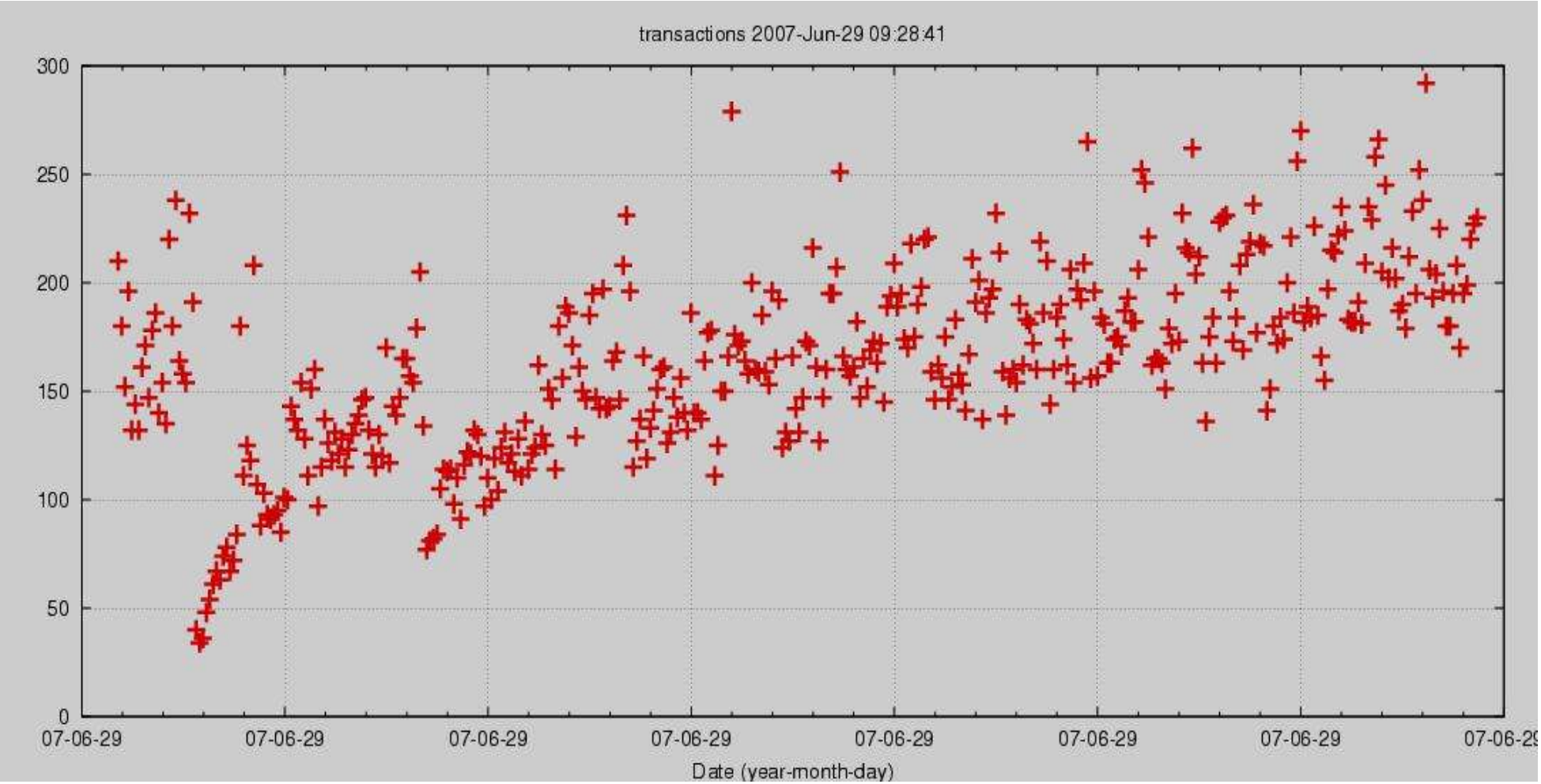
pnfs transaction rates for raid via FC switch



one pnfs server on BlueArc



two pnfs servers on BlueArc



Moving services (HW)

We currently do not plan automatic service failover.

On the node which is accepting new service manually:

1. Make changes to zone map on switch to allow this host to see broken hosts raid and disallow the old host.

2. Make changes to fstab to allow this node to mount broken hosts raid.

3. Mount broken nodes raid and check it's integrity.

4. Run script to start broken nodes services on this node.

On the broken node when it is repaired.

1. Boot up in mode that allows you to change fstab to not mount raid.

2. This node now becomes spare.

Moving Services (SW)

- Enstore services are moved by manually changing enstore configuration and stopping / starting service
- Some services will use a network service address (SA), when SA is assigned to a service rather that to a host. This is needed because enstore clients use predefined host address to access these services (pnfs, web). The anchor SA will be handled in collaboration with network group.
  - Anchor SA will be assigned to service aliases by network group.
  - Configure servers to use SA and aliases.
  - An SA reassignment script to execute ifconfig up / down to start a new alias IF
- The following services to use SA: pnfs, web server, enstore configuration server

System Installation

Enstore runs on tens of nodes.

- use kickstart disk to install OS. Kick start disks are created for the new OS release and modified as needed. A separate kickstarts for enstore servers and movers.
- configuration and synchronization of system configuration files provided by cfengine running on a dedicated host serving also as yum and ftp server for enstore and related products distribution.
- we consider RPM installation and configuration of enstore and related products and services.
- enstore product updates will still be done via CVS. Major releases will be (eventually) done via RPM

Software

- Current enstore SW was not flexible enough to meet the goals of the project. Some changes were done to make enstore components completely configurable. This allows services to move by changing the system configuration.
- The necessary changes were made in the enstore configuration file. The modifications to the code were made to use new configuration parameters.
- RPMs were created to install and configure enstore, ftt, Python-enstore, web server configuration
- Cron jobs are coupled now with enstore servers. They are grouped in corresponding files in /etc/cron.d.

Milestones

- Completion of requirements, design and breakdown into implementation – February 2007.
- Purchase of HW – 2006.
- Installation of hardware - June 2007.
- beta software SDE, testing and bug fix cycle – July 2007.
- Service Address implementation – August 2007
- Documentation – September 2007.
- Deployment of 1st system (reconfiguring hardware and installing software. September 2007. (see next page)

Deployment Plan

- Provide HW environment, SSA Group. (1st instance – June).
- Develop S/W environment. DMD (August).
- Test deployment. (June – August).(http://d0ensrv2n.fnal.gov/enstore/)
- Iterate 1-3 Until final production.
- Deploy to production.

  a) d0en – Sept
  b) cdfen – Oct
  c) stken – Nov*

  * 1 scheduled downtime per month

  d0en and cdfen will require extended downtime to copy DBs from old raids to new.

Staffing

- SSA provides (planned: .75 FTE, real: .5).
- DMD provides (planned: 2 FTE, real:1.25).
- Staff is being drawn from this project by other projects. This has extended the time required to complete the project.

Conclusion

- SDE allows reduction of stress and load on administrators and offload hardware maintenance to vendor.
- Reduction of downtime associated with hardware failures of servers and their reconfiguration.
- Better rolling upgrades.
- Reduces the impact to users.
- Allows redirection of effort to other enstore and dcache activities.
- Better packaging for external deployment (as byproduct).