Open Science Grid

Frank Würthwein
OSG Application Coordinator
OSG Extension Lead
Experimental Elementary Particle Physics
UCSD

Overview

- ⇒ OSG in a nutshell
- ⇒ Using the OSG
- ⇒ Present Utilization & Expected Growth
- ⇒ Near term Future

OSG in a nutshell

- High Throughput Computing
  - Opportunistic scavenging on cheap hardware.
  - Owner controlled policies.
- "open consortium"
  - Add OSG project to an open consortium to provide cohesion and sustainability.
- Heterogeneous Middleware stack
  - Minimal site requirements & optional services
  - Production grid allows coexistence of multiple OSG releases.
- "Linux rules": mostly RHEL3/4 on Intel/AMD
- Grid of clusters
  - Compute & storage (mostly) on private Gb/s LANs.
  - Some sites with (multiple) 10Gb/s WAN "uplink".

Using the OSG

Authentication & Authorization
Moving & Storing Data
Submitting jobs & “workloads”
Strategies for Success

Making the Grid attractive

⇒ **Minimize entry threshold for resource owners**
- Minimize software stack.
- Minimize support load.

⇒ **Minimize entry threshold for users**
- Feature rich software stack.
- Excellent user support.

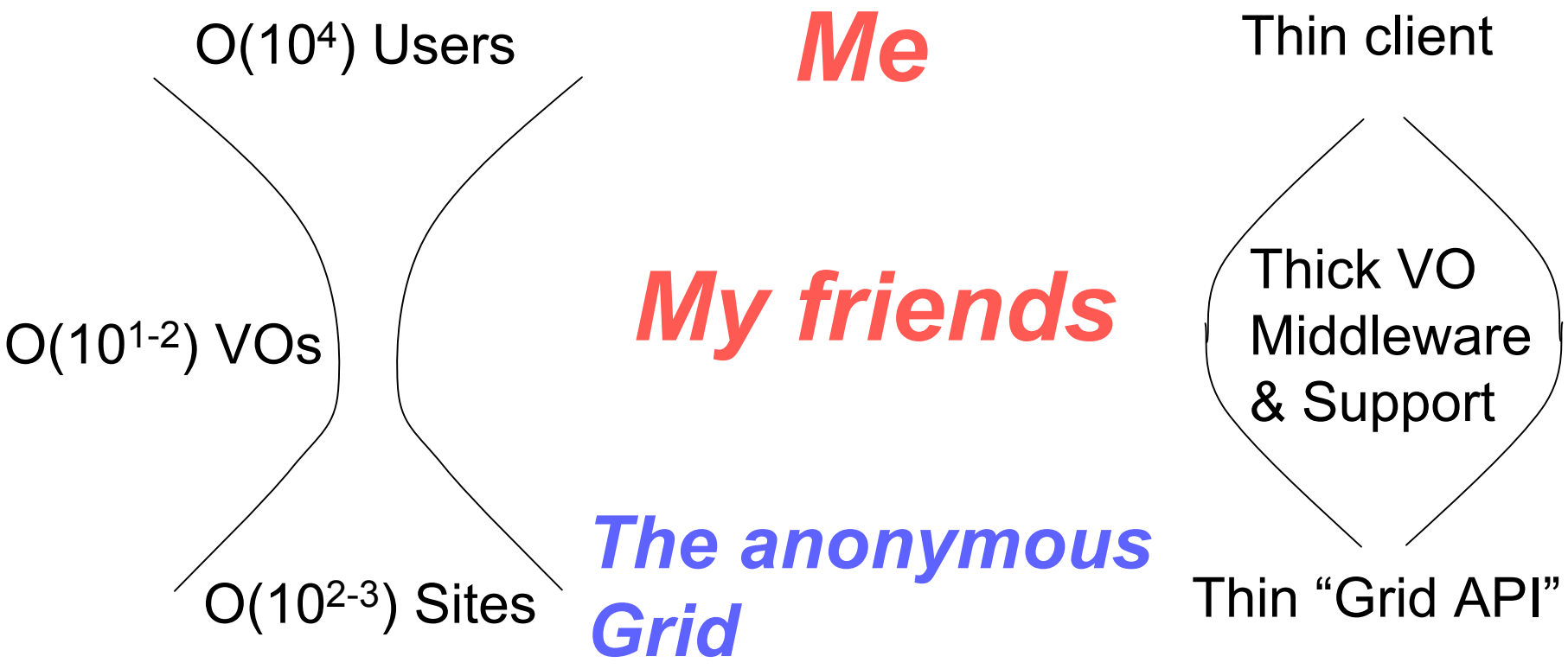
Resolve contradiction via "thick" Virtual Organization layer of services between users and the grid.

(Talks in "Enabling User Communities" Session Monday afternoon.)

Me -- My friends -- The grid

Domain science specific

Common to all sciences



Authentication & Authorization

- **OSG Responsibilities**
  - X509 based middleware
  - Accounts may be dynamic/static, shared/FQAN-specific
- **VO Responsibilities**
  - Instantiate VOMS
  - Register users & define/manage their roles
- **Site Responsibilities**
  - Choose security model (what accounts are supported)
  - Choose VOs to allow
  - Default accept of all users in VO but individuals or groups within VO can be denied.

User Management

- ⇒ *User obtains DN* from CA that is vetted by TAGPMA
- ⇒ *User registers with VO* and is added to VOMS of VO.
  - VO responsible for registration of VOMS with OSG GOC.
  - VO responsible for users to sign AUP.
  - VO responsible for VOMS operations.
    - VOMS shared for ops on multiple grids globally by some VOs.
    - *Default OSG VO exists for new communities & single PIs.*
- ⇒ *Sites decide which VOs to support* (striving for default admit)
  - Site populates GUMS daily from VOMSes of all VOs
  - Site chooses uid policy for each VO & role
    - Dynamic vs static vs group accounts
- ⇒ User uses whatever services the VO provides in support of users
  - *VOs generally hide grid behind portal*
- ⇒ Any and all *support is responsibility of VO*
  - Helping its users
  - Responding to complains from grid sites about its users.

Moving & storing data

- **OSG Responsibilities**
  - Define storage types & their APIs from WAN & LAN
  - Define information schema for "finding" storage
  - All storage is local to site - no global filesystem!
- **VO Responsibilities**
  - Manage data transfer & catalogues
- **Site Responsibilities**
  - Choose storage type to support & how much
  - Implement storage type according to OSG rules
  - *Truth in advertisement*

Disk areas in some detail:

- ⇒ Shared filesystem as *applications area* at site.
  - Read only from compute cluster.
  - Role based installation via GRAM.
- ⇒ Batch slot specific *local work space*.
  - No persistency beyond batch slot lease.
  - Not shared across batch slots.
  - Read & write access (of course).
- ⇒ SRM/gftp controlled *data area*.
  - "persistent" data store beyond job boundaries.
  - Job related stage in/out.
  - SRM v1.1 today.
  - SRM v2.2 expected in Q2 2007 (space reservation).

Submitting jobs/workloads

- **OSG Responsibilities**
  - Define Interface to batch system
  - Define information schema
  - Provide middleware that implements the above.
- **VO Responsibilities**
  - Manage submissions & workflows
  - VO controlled workload management system or wms from other grids, e.g. EGEE/LCG.
- **Site Responsibilities**
  - Choose batch system
  - Configure interface according to OSG rules
  - *Truth in advertisement*

Applications & Runtime Model

- User specific portion that is small and comes with the job.
- VO specific portion that is large and is preinstalled.

- CPU access policies vary from site to site
  - Ideal runtime ~ O(hours)
    - Small enough to not loose too much due to preemption policies.
    - Large enough to be efficient despite long scheduling times of grid middleware.

Simple Workflow

- **Install Application Software at site(s)**
  - VO admin install via GRAM.
  - VO users have read only access from batch slots.
- **"Download" data to site(s)**
  - VO admin move data via SRM/gftp.
  - VO users have read only access from batch slots.
- **Submit job(s) to site(s)**
  - VO users submit job(s)/DAG via condor-g.
  - Jobs run in batch slots, writing output to local disk.
  - Jobs copy output from local disk to SRM/gftp data area.
- **Collect output from site(s)**
  - VO users collect output from site(s) via SRM/gftp as part of DAG.

Late binding

Talks by:
Maeno: Monday afternoon
Sfiligoi, Padhi: Tuesday afternoon

- Grid is a hostile environment:
  - Scheduling policies are unpredictable
    - Many sites preempt, and only idle resources are free
  - Inherent diversity of Linux variants
  - Not everybody is truthful in their advertisement
- Submit "pilot" jobs instead of user jobs
- Bind user to pilot only after batch slot at a site is successfully leased, and "sanity checked".
- Re-bind user jobs to new pilot upon failure.

Status of Utilization

OSG job = job submitted via OSG CE

“Accounting” of OSG jobs not (yet) required!

Grid of sites

- IT Departments at Universities & National Labs make their hardware resources available via OSG interfaces.
  - CE: (modified) pre-ws GRAM
  - SE: SRM for large volume, gftp & (N)FS for small volume
- **Today's scale:**
  - 20-50 "active" sites (depending on definition of "active")
  - ~ 5000 batch slots
  - ~ 1000TB storage
  - ~ 10-15 "active" sites with shared 10Gbps or better connectivity
- **Expected Scale for End of 2008**
  - ~50-100 "active" sites
  - ~30-50,000 batch slots
  - Few PB of storage
  - ~ 25-50% of sites with shared 10Gbps or better connectivity

OSG use by Numbers

39 Virtual Communities

6 with >1000 jobs max.

(5 particle physics & 1 campus grid)

4 with 500-1000 max.

(two outside physics)

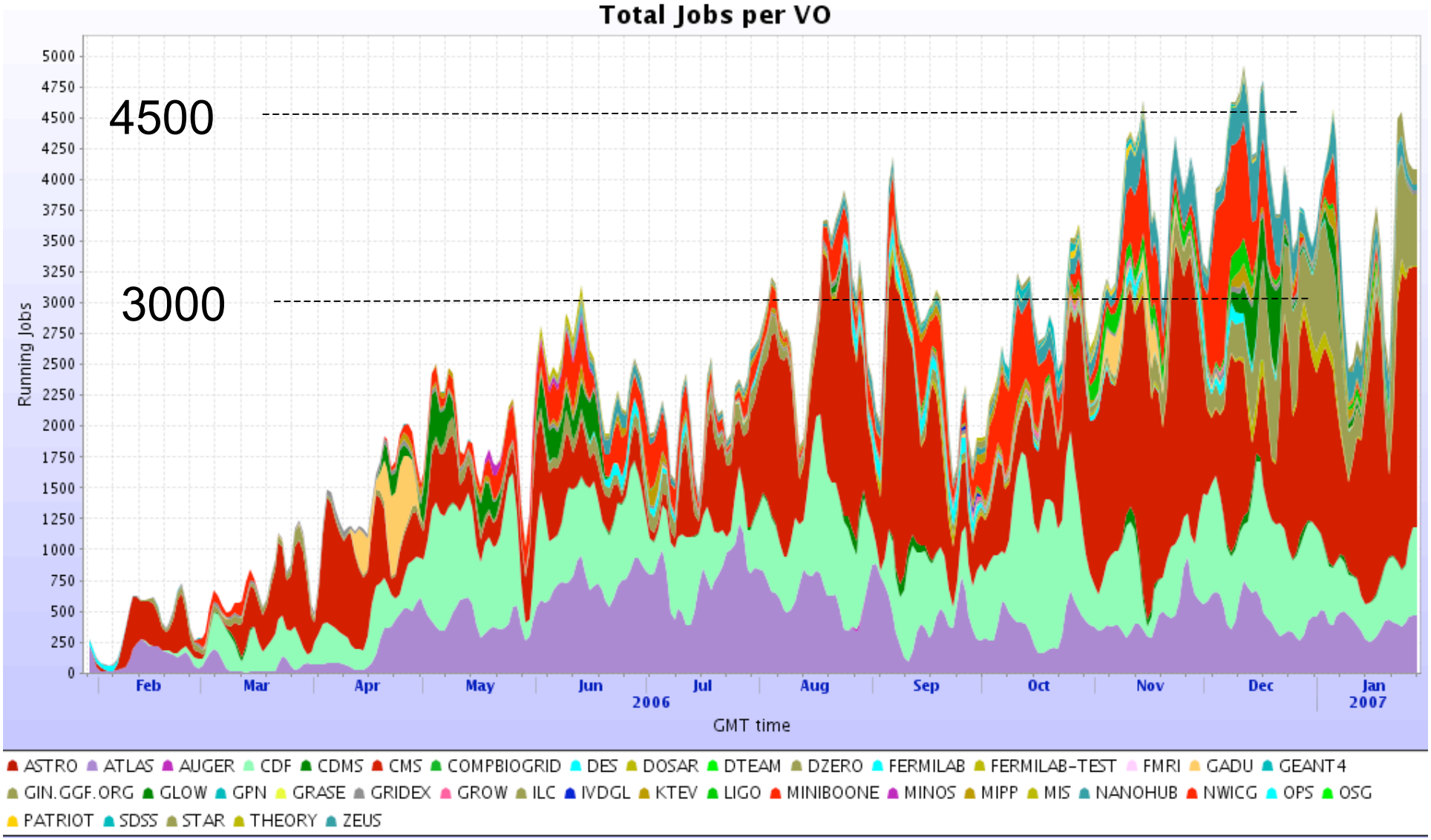
10 with 100-500 max

(campus grids and physics)

Running Jobs

Farm	Last value	Min	Avg	Max
ASTRO	0	0	0	0
ATLAS	463	0	420.8	1555
AUGER	0	0	0.334	64
CDF	734	0	514.5	2086
CDMS	0	0	11.57	880
CMS	2021	0	791.1	3719
COMPBIOGRID	0	0	0.345	10
DES	0	0	1.486	62
DOSAR	9	0	20.74	226
DTEAM	0	0	0.183	2
DZERO	572	0	135.6	1825
FERMILAB	0	0	24.43	562
FERMILAB-TEST	0	0	0.036	1
FMRI	0	0	0	0
GADU	0	0	29.61	754
GEANT4	0	0	0	2
GIN.GGF.ORG	0	0	0.007	4
GLOW	4	0	45.65	1313
GPN	0	0	0	0
GRASE	0	0	0.301	14
GRIDEX	33	0	25.93	268
GROW	0	0	2.693	110
ILC	0	0	0	0
IVDGL	0	0	0.714	73
KTEV	0	0	22.12	288
LIGO	0	0	23.1	369
MINIBOONE	0	0	183.4	2000
MINOS	0	0	5.829	170
MIPP	0	0	13.52	208
MIS	0	0	0.444	71
NANOHUB	37	0	81.92	600
NWICG	0	0	0	2
OPS	0	0	0.011	4
OSG	0	0	0.316	27
PATRIOT	0	0	3.477	194
SDSS	4	0	8.41	197
STAR	128	0	21.72	334
THEORY	0	0	5.267	73
ZEUS	0	0	2.721	205
Total	**4005**		**2398**	

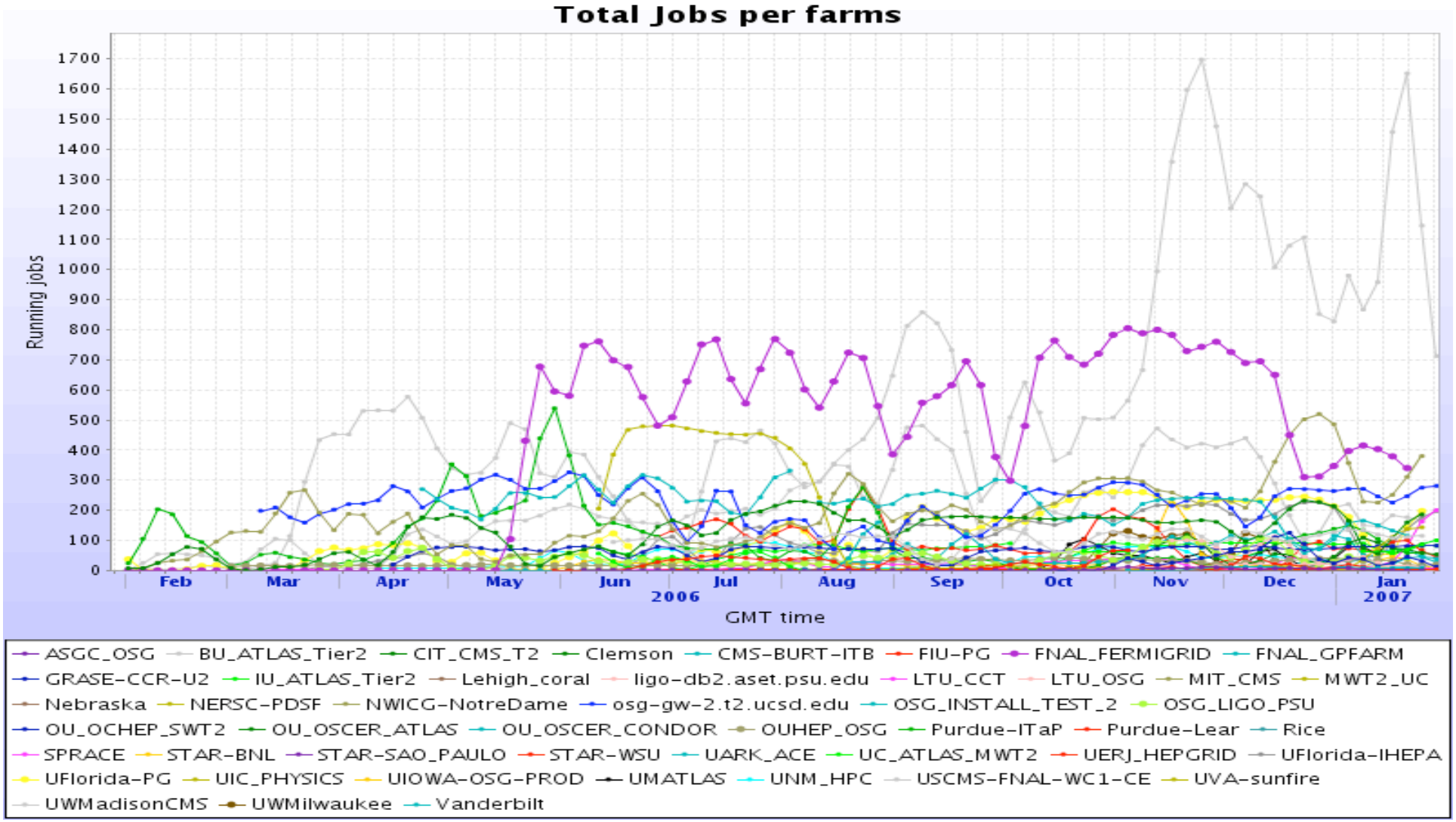
Number of running (and monitored) “OSG jobs” within last year.



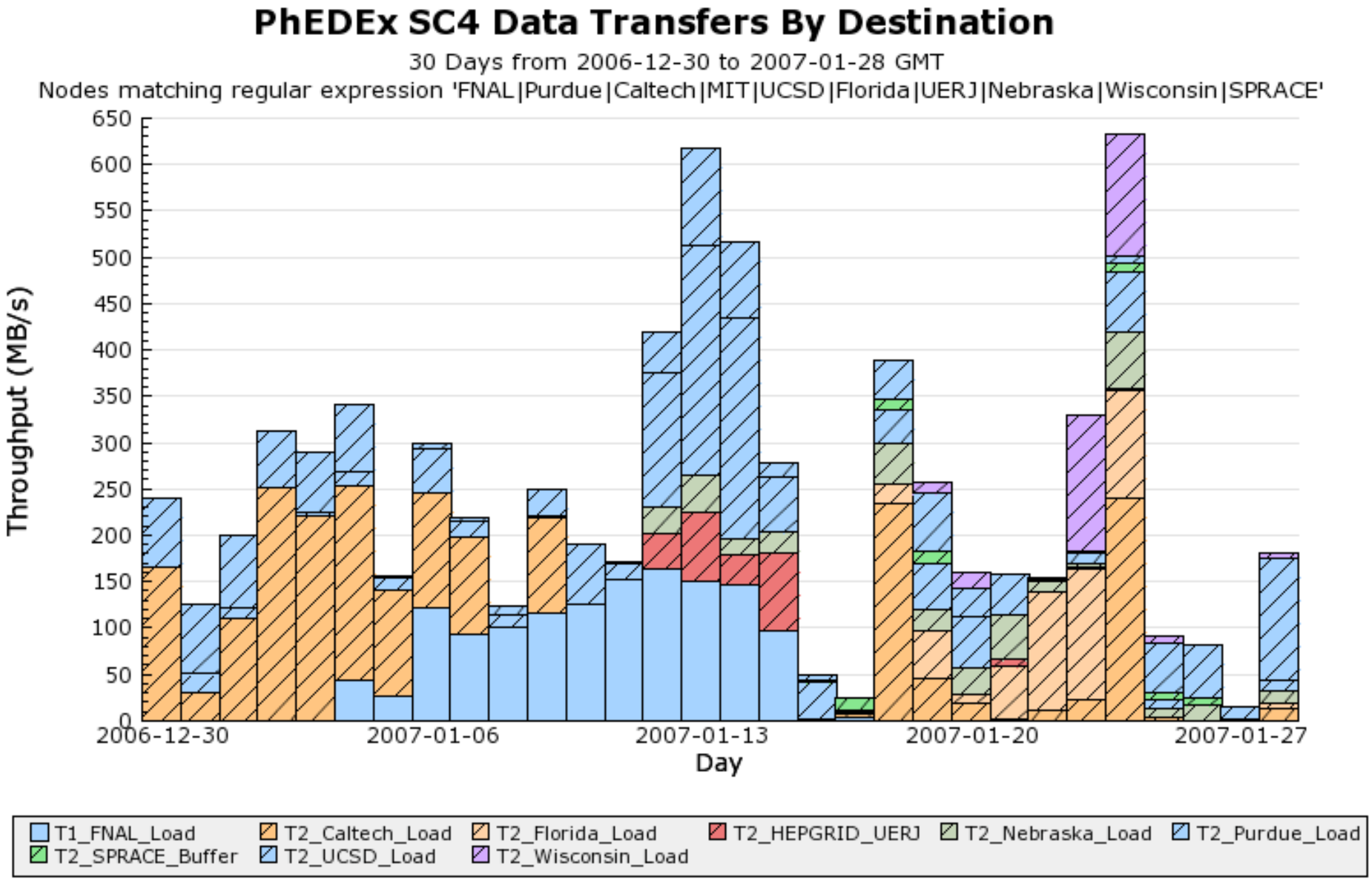
Number of jobs running at sites:

Many small sites, or with mostly local activity.

>1k max	5 sites
>0.5k max	10 sites
>100 max	29 sites
Total:	47 sites



CMS on OSG January 2007



Next Steps in OSG facility: Dotting the i's and crossing the t's

- Focus on Accounting
  - OSG 0.6 comes with first mandatory accounting system
  - Wall clock time, data transfers, space utilization are accounted for.
- Focus on large scale managed storage
  - Added SRM/dCache 1.7 to OSG 0.6
  - SRM/dCache 1.8 coming with space reservation as OSG 0.6.x
- Focus on Information System
  - CEMon @ sites and centralized OSG infosys @ GOC
  - Truth in advertisement
    - Task force on GIP attributes, including site validation & ticketing

For more, see site validation session on Monday, and "Effectiveness session on Tuesday.