

Professional usage of the VL-e/BIG GRID infrastructure

An IBM Life Science use-case
and the lessons learned

Gridforum.nl 2008 Business Day

Alex de Landgraaf <alex@aperte.nl>

IBM - Vrije Universiteit Amsterdam

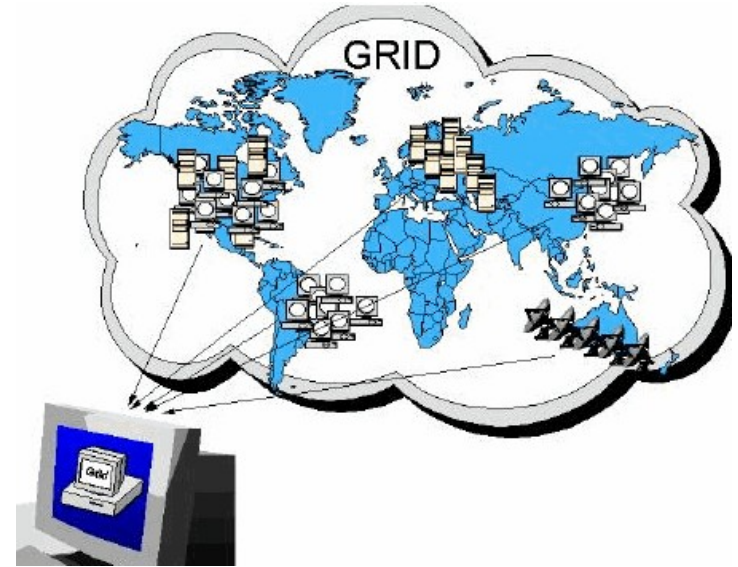


How I hijacked this Business Day

- Intern at IBM on VL-e/grids
- Paper for IBM: Grids and Web 2.0
- Internship goal:
 - Theoretical and practical evaluation of the use of the dutch grid infrastructure for production environments
- Specifically: area of Healthcare & Life Sciences, data grid only
- Disclaimer: These are my own opinions, IBM and VU are not to blame

Why evaluate grids?

- High performance
- Secure
- Scalable
- Interoperability
- Also: nearly-there research technology for quite a few years, in spite of considerable funding
- Commercial/production usage seen as a goal. How far are we?



VL-e/BIG GRID infrastructure

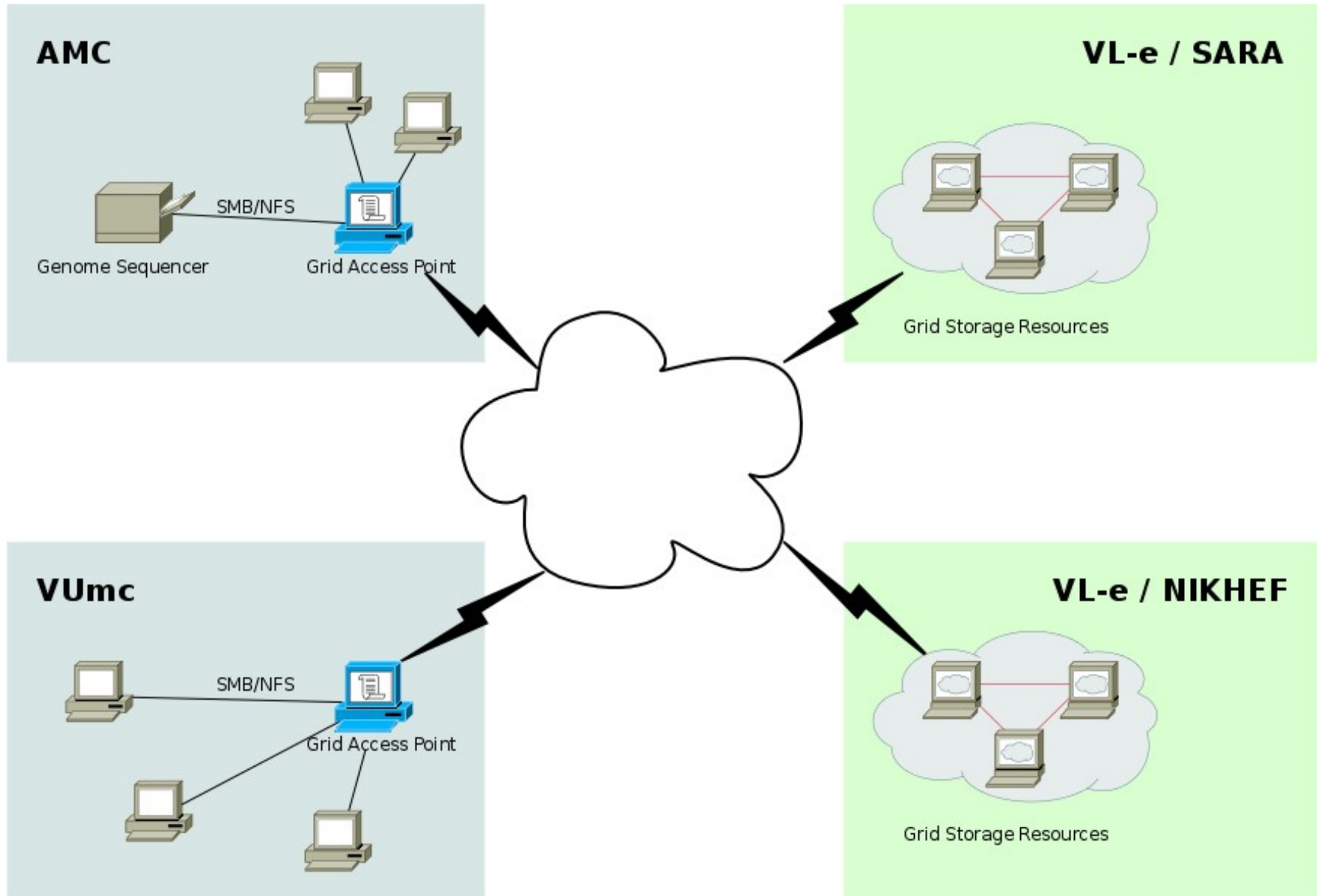
- “Virtual laboratory for e-science”
- Provide a generic software platform for research.
- Used VL-e Proof-of-Concept grid
- BIG GRID – sustainable nation-wide grid infrastructure, extension of VL-e PoC
- EU-initiated EGEE project: 50k CPUs, 15PB, makes use of gLite middleware

AMC case

- Genome sequencer FLX, Roche
- Storing and exchange of raw data required, 10TB/year
- 2 runs/day, EUR 5-7k/run
- Sensitive data
- Used by non IT specialists, both research and clinical, on multiple UMC sites
- Necessary to integrate grid storage into existing infrastructure

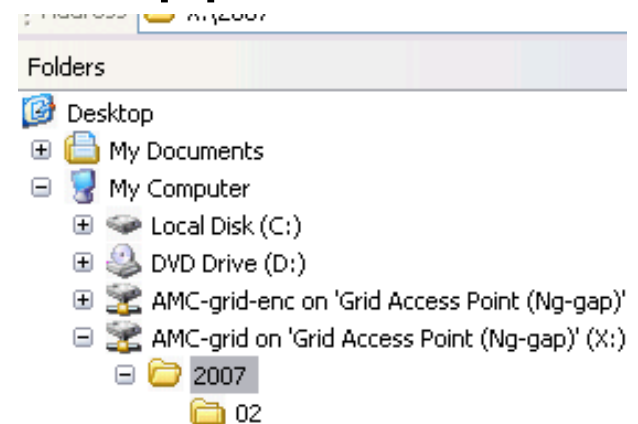


Overview of the solution



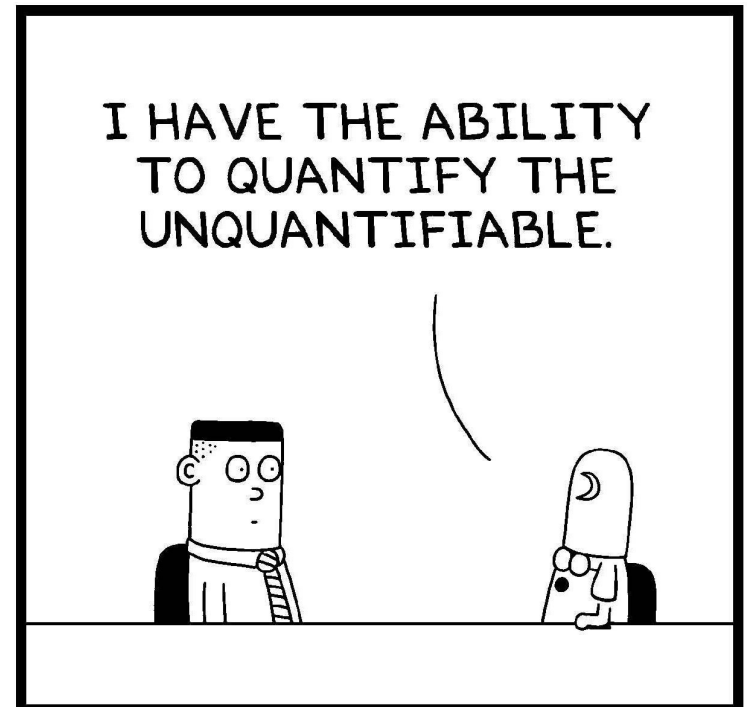
Pilot implementation details

- “Grid-access-point” server at AMC
- Developed a grid-filesystem interface (VGFS) Linux FUSE and gLite APIs
- Files on the grid accessible via normal network protocols, all OS, existing (IBM) products
- Users access grid data resources like any NAS, no need for training or specialized support
- Optional on-the-fly encryption



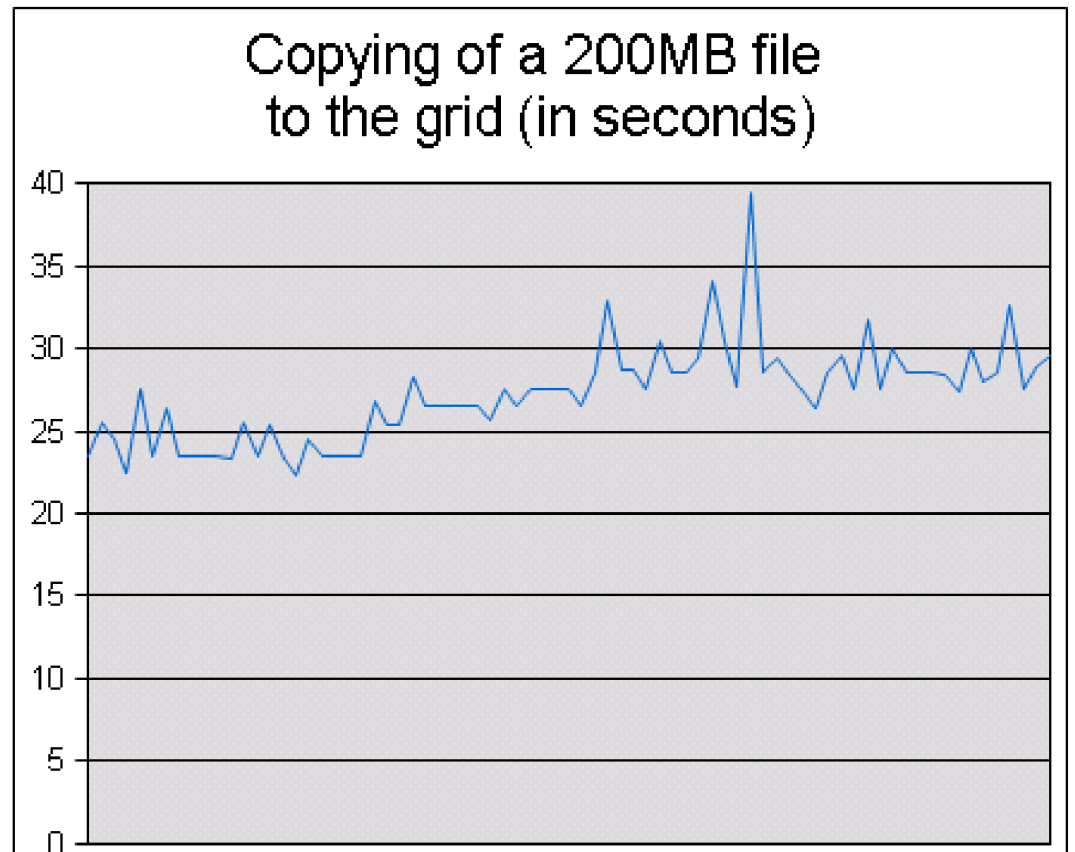
Assessment of solution

- Security ✓
- Privacy ✓
- Usability ✓
- Performance ?
- Availability ?



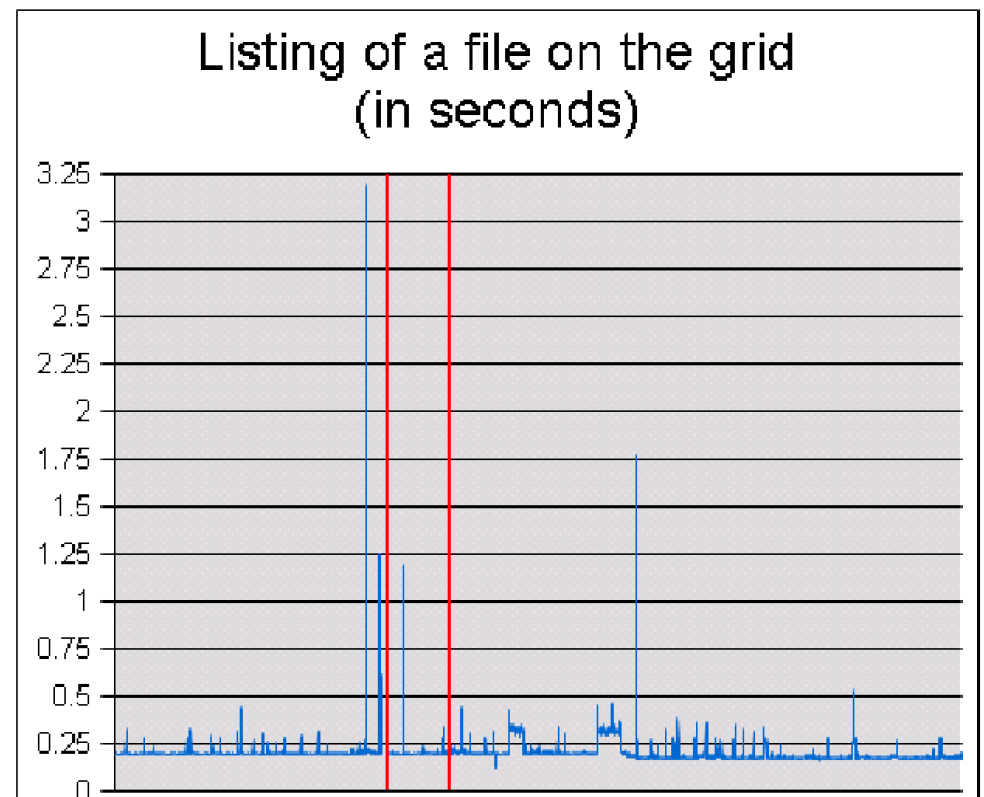
Performance

- How do we test this?
- Simply copy a file a few hundred times
- VGFS-to-grid:
 - 7.4MB/s,
 - includes latency
- Sufficient for case,
 - 14GB \approx 1/2 hour



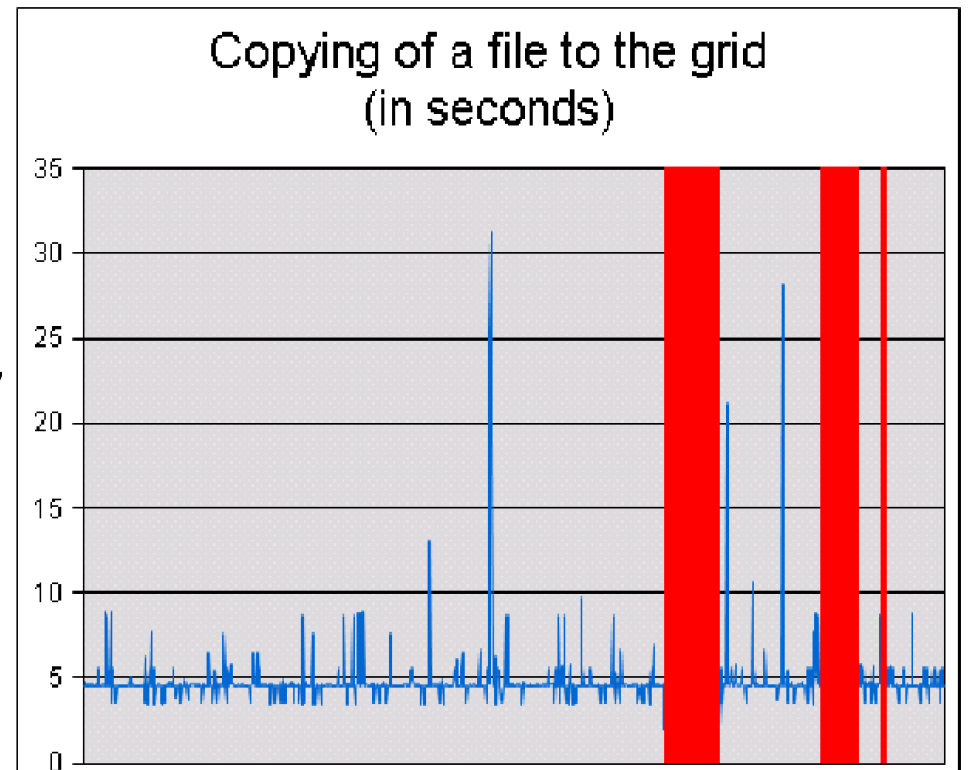
Availability (1/2)

- How do we test this?
- Over 4 weeks, list and copy operations, use gLite utilities directly
- List file:
 - 99.93% uptime
 - 0.2 second latency
- Acceptable



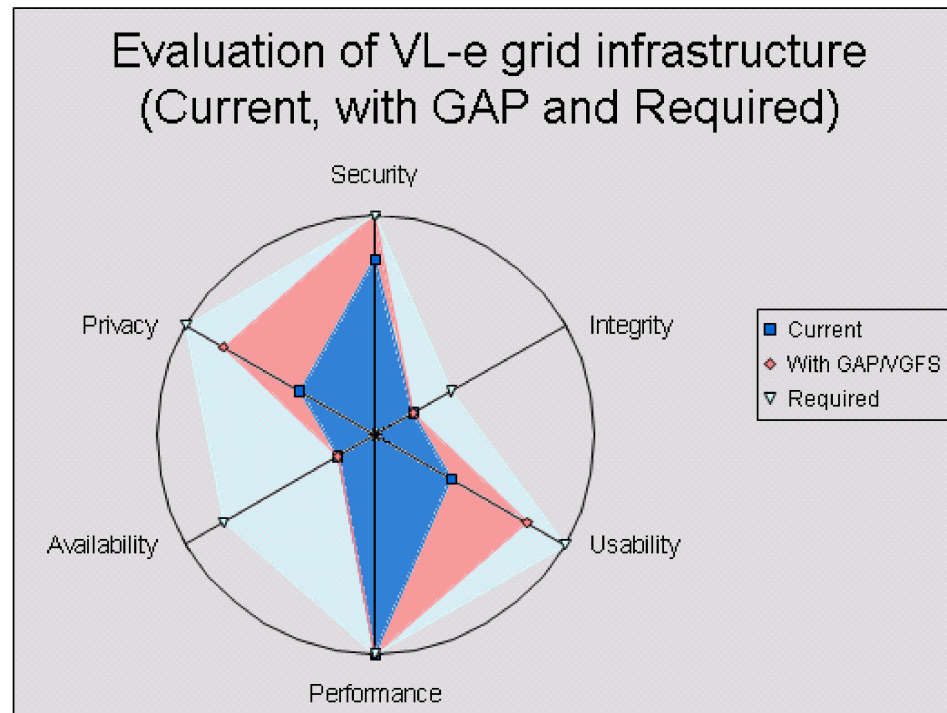
Availability (2/2)

- Copy file (100KB):
 - 93.34% uptime (1 in ~ 12 failed)
 - 4.5 second latency
- Not acceptable
- Likely reasons:
 - Single points of failure,
 - lack of redundancy in
 - key grid services



Conclusion

- Able to bring grid usability to acceptable levels for AMC case
- VL-e PoC good enough for eScience, not yet for Business
- Availability is critical for production environments, too often overlooked in discussions about grids (too boring?)
- Reliability is “Achilles heel” for off-site resources and (Web 2.0) services: more things can (and will) go wrong



Questions?

- Grids for eScience?
 - Grids for Business?
 - Grids for Everyone?
-
- Full non-IBM report, executive summary:
Ask me for dead-tree version or
alex@aperte.nl for PDF