



# Life Sciences & The Dutch Grid

**An Analysis from a Grid Supporter's perspective**

***IWPLS'09, Edinburgh***

***Evert Lammerts***

***SARA Computing and Networking Services***

***High Performance Computing & Visualization***

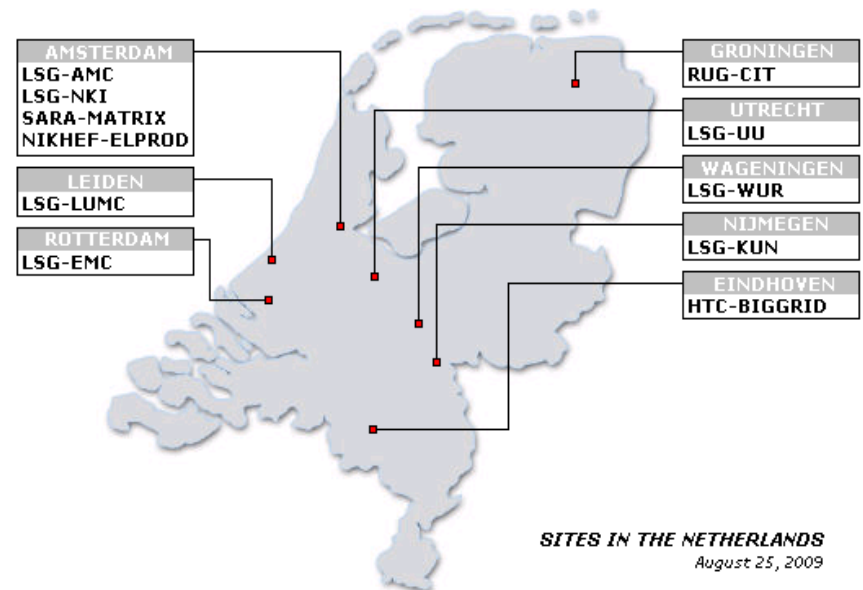
***e-Science Support***

**evert.lammerts@sara.nl**



# Life Sciences In The Netherlands: 11 Sites

- ▲ **LSG-EMC:**
  - ▲ 32 cores (32 bit)
  - ▲ 20 TB storage
- ▲ **6 Other LSG-\* sites:**
  - ▲ 16 cores (32 bit)
  - ▲ 1.5 TB storage
- ▲ **SARA-MATRIX**
  - ▲ 808 cores (32 bit)
  - ▲ 741 TB storage
- ▲ **NIKHEF-ELPROD**
  - ▲ 1144 cores
- ▲ **RUG-CIT**
  - ▲ 172 cores
- ▲ **HTC-BIGGRID**
  - ▲ 1640 cores



# Context & Organization

## ■ *BigGrid* project

- Aims to provide and maintain the national Grid Infrastructure
- Founding partners: Netherlands Bioinformatics Center (NBiC), Netherlands Computer Facilities Foundation (NCF) and the Netherlands Institute for Subatomic Physics (Nikhef)
- Core partners: SARA, Philips Research, University of Groningen

## ■ *Life Science Grid* project: all LSG-\* sites

- Initiated by SARA
- Commissioned by NCF & NBiC

## ■ Updates:

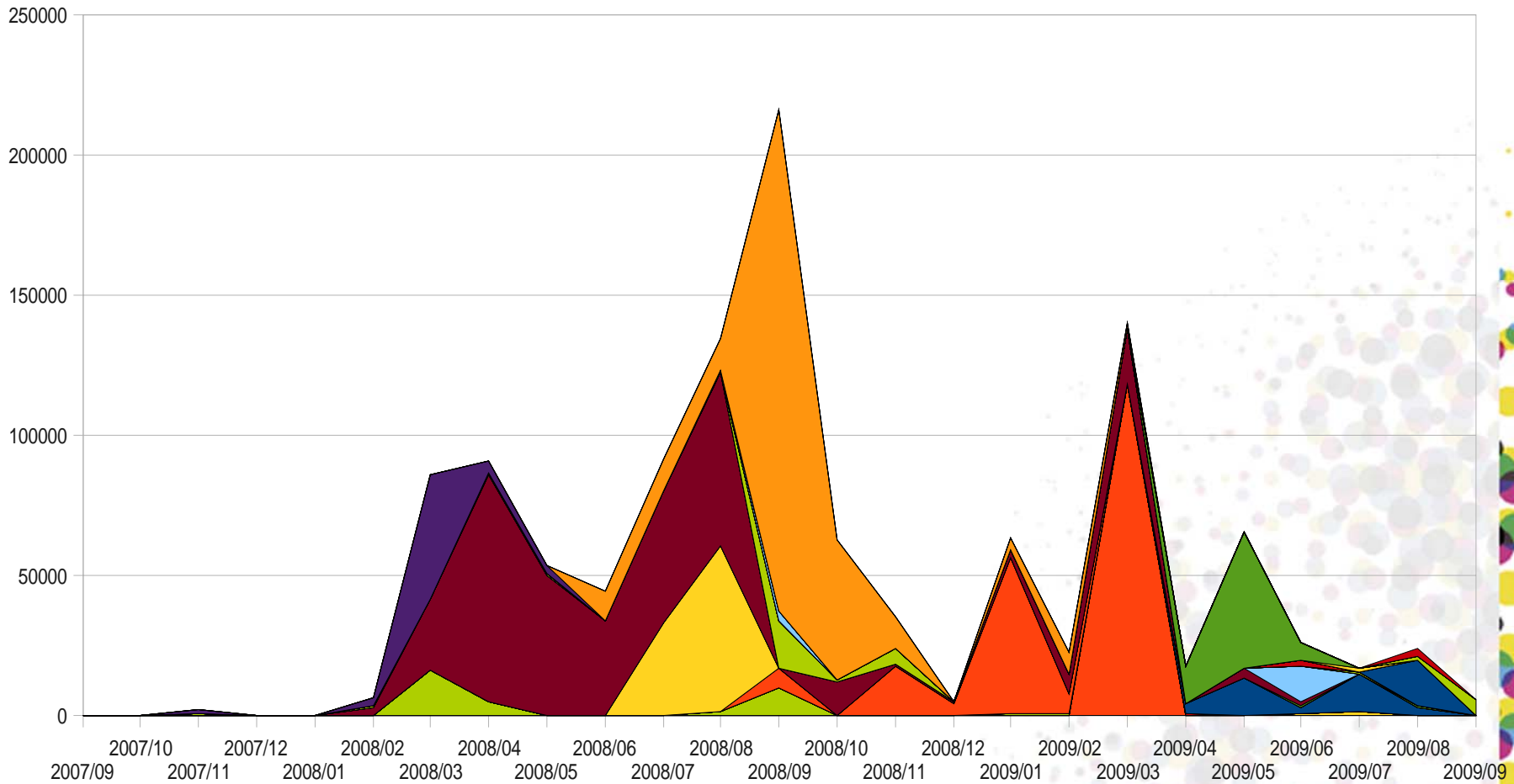
- LSG-\* to 64 bit and 20 TB storage
- SARA-MATRIX to 2216 cores and 3,4 Peta Byte storage
- Other sites will also upgrade

# What?

- ▶ **Statistical usage of the Grid**
- ▶ **Suggestions for analysis**
- ▶ **A first try:**
  - ▶ **The types of problems**
  - ▶ **The origin of these problems**
  - ▶ **Possible solutions**



# Statistics (usage)



# Statistics (data & context)

## Context:

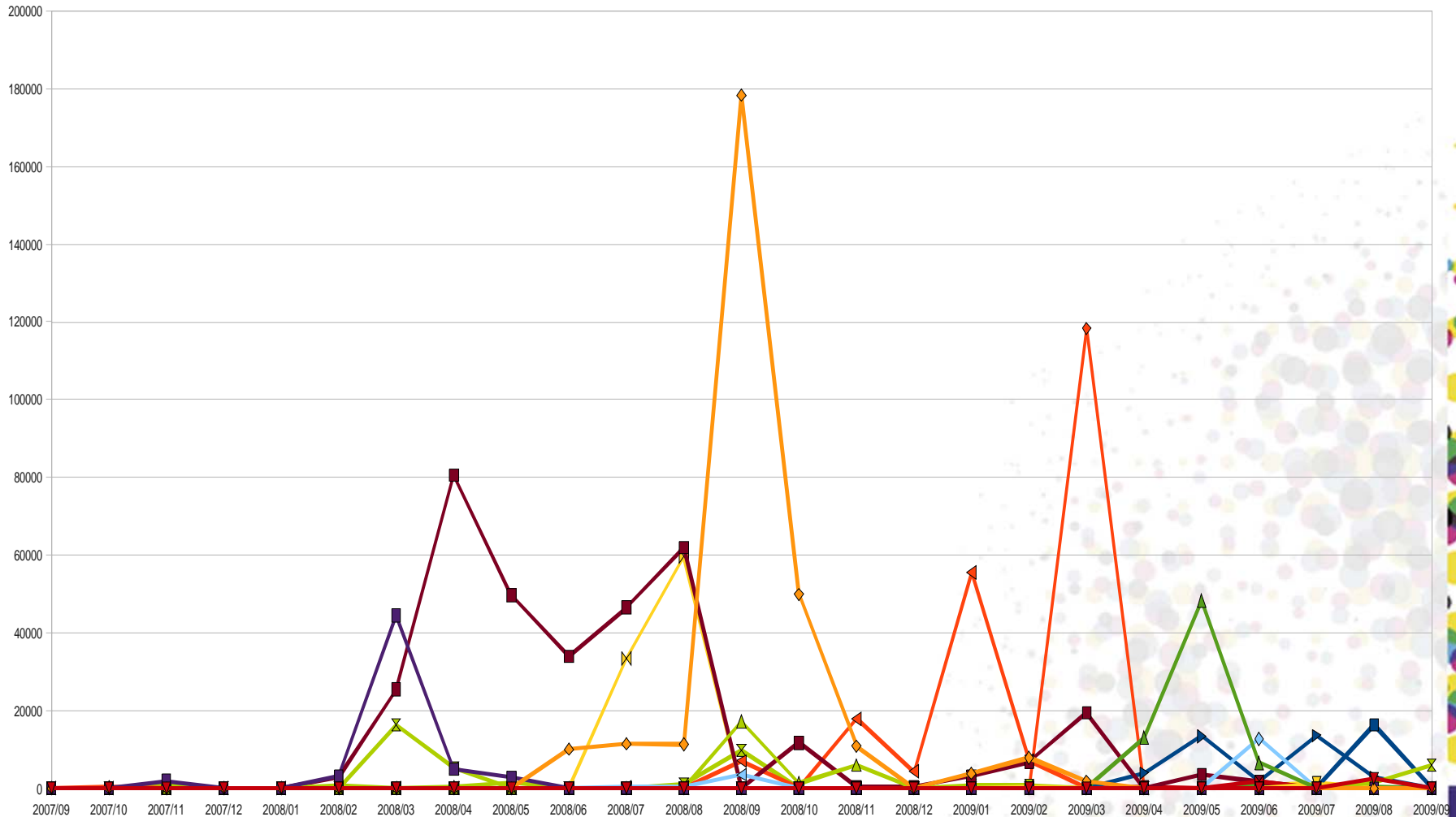
- Only LSGrid VO
- Only clusters administered by SARA, excluded are:
  - ▶ Nikhef
  - ▶ Philips (not fully functional)
  - ▶ Groningen University (problems with SRM)
- No comparisons:
  - ▶ How does Life Science perform on our national cluster LISA?
  - ▶ How does this compare to other countries?

## LSGrid – the largest Life Science VO of the Netherlands:

- 67 registered Distinguished Names
- 35 'active' users (at least 1 month => 3 minutes)
- 7 major users (at least 1 month => 20,000 hours), of which...
  - ▶ 1 is a SARA employee
  - ▶ 4 whose work is intensely supported by SARA
  - ▶ 2 whose work is indirectly supported by SARA

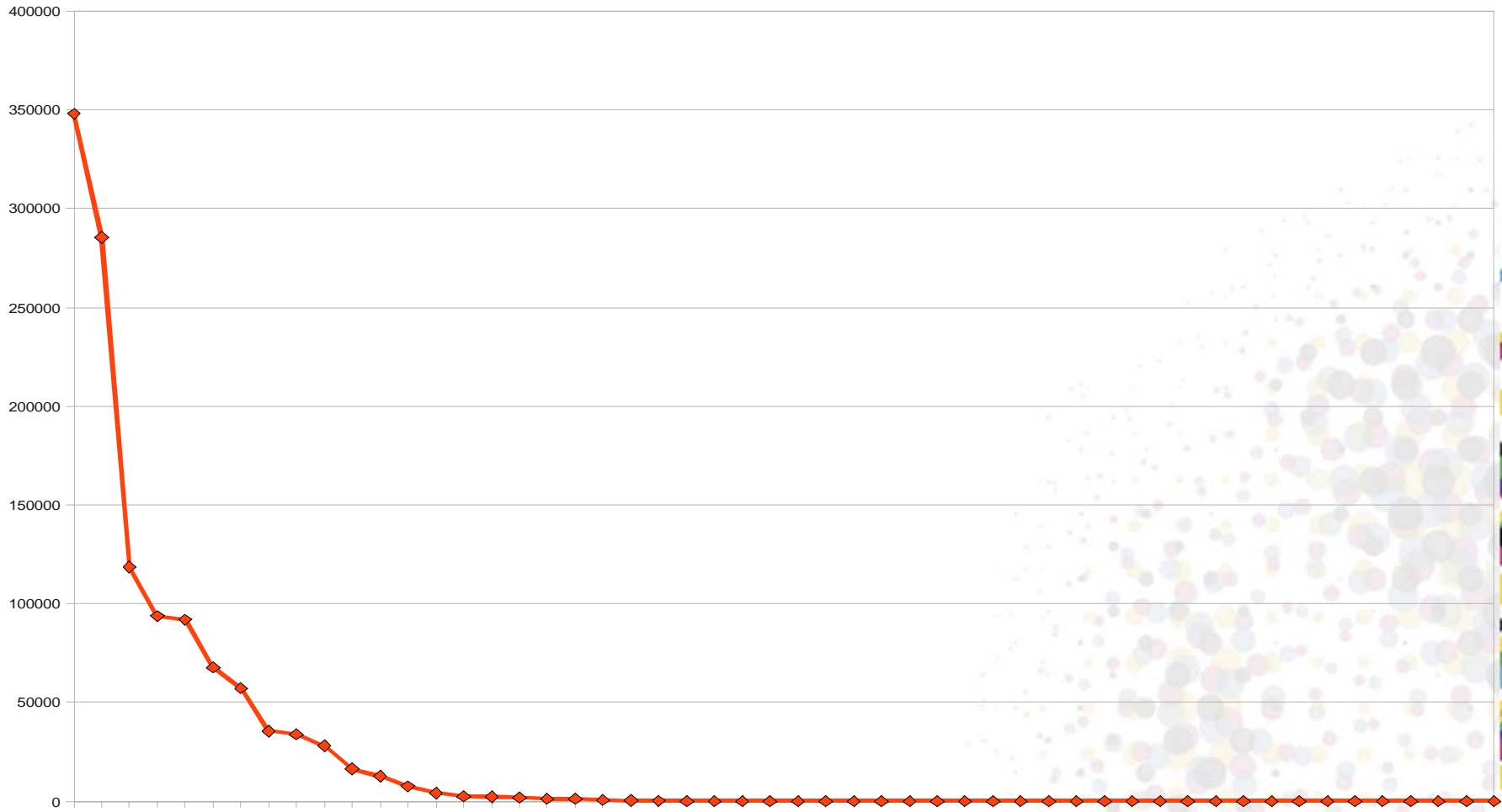


# Statistics (the 7 users)





# Statistics (persistence)



# These Statistics Imply...

- ▶ **That very little of the LSGrid VO members manage to use the Grid to an extent that indicates it is useful for their purpose**
- ▶ **That the members that do manage are the ones that need a significant amount of support (and are therefore probably better at networking and communicating than they are at using the system...)**
- ▶ **That, from the perspective of the Life Scientist, the amount of effort needed to work with the Grid is not proportional to its usefulness**

# Resulting Questions

- **Wat is the cause of these statistics?**
  - **Do we our jobs as well as we should?**
    - ▶ **The quality of our support / documentation**
  - **Are the systems flawed?**
    - ▶ **The concepts of distributed computing**
  - **Is the software or are its interfaces flawed?**
    - ▶ **The middleware**
    - ▶ **The tools**
  - **Is the Life Scientist not capable?**
    - ▶ **Knowledge of technology**

# How To Answer Them?

- **How can we analyze the problems implied by the statistics?**
  - **Better monitoring / accounting:**
    - ▶ **Usage of the systems! (Cesga is not good enough)**
    - ▶ **Support Key Performance Indicators (KPI's)!**
  - **Philosophy of Science:**
    - ▶ **Do our systems and interfaces comply with the assumptions and foundations of Life Science?**
  - **Investing more effort in analytics and less on ad-hoc solutions**

# A first try: Types of problems

- **Organizational: We *don't* do our jobs as well as we should!**
  - Communication, both internal and external, is not sufficient
  - No ways to enforce appointments
  - No KPI's, so no self reflection
  
- **Technical: The software *is* flawed!**
  - Flaws in the middleware
  - Flaws in the interfaces
  - Little effort to provide interfaces *specific* to the LS – generic is still the keyword
  
- **Naïve (– the positive type): The Life Scientist is *not* always capable to use the Grid**
  - A lack of understanding of the concepts of distributed computing
  - An extremely steep learning curve

**BUT: are scientists from other disciplines more successful?**

# Naivety

- ▶ **Life Sciences: A shift from one research modality to the other:**
  - ▶ Hypothesis driven
  - ▶ Data driven
  
- ▶ **Data driven research requires a different Life Scientist:**
  - ▶ An understanding of the concepts of data mining
  - ▶ An understanding of systems capable of efficient data mining
  
- ▶ **We notice the same with different interfaces (Diane, Monteur, Taverna, ...):**
  - ▶ Concepts of workflows are not understood
  - ▶ Concepts of pilot jobs are not understood
  - ▶ Etc. etc. etc.

***DOES THE DATA DRIVEN LIFE SCIENTIST EXIST?  
(.. or at least, in the form we assume he does?)***

# Technical

## ■ Software:

- Flaws in middleware
- Flaws in interfaces

## ■ Distributed computing:

- The growth of the Grid, in hard- and software as well as involved people, is parallel to the growth of the surface on which errors can occur;  
*How often do updates result in tickets related to a dependency of the updated component?*

## ■ Tailored to a different use-case!

- Try putting X-thousand small files (Life Scientists do this!) on an the Grid with SRM
- Technology push

# Organizational

## Internal organization

- Updates, upgrades, and maintenance are not well communicated internally nor externally - EGEE Intervention procedures are a good start but we need more
- Support is not evaluated!

## External organization

- Who is responsible for what? Grid Support, VO manager, who do I need for what?
- SPoE to Grid documentation and information for the Life Scientist stands for Sixhundred Points of Entry – too many different systems

# Solutions (?)

- ▶ Shift from hypothesis- to data driven = desktop to large scale: *so bring the large scale to the desktop*
- ▶ But not *generic*: replacing one complexity with the other has not worked
- ▶ Good potential:
  - ▶ Annotated application services
  - ▶ Customized workflows
- ▶ Data Storage: WebDAV?
- ▶ Middleware: Virtualization, cloud?
- ▶ Evaluation of software and support!

# Solutions (?)

- ▶ Reduce 600 points of entry to a single point of entry
- ▶ Make VO's stronger



# Thanks for your attention!

▶ Evert Lammerts  
SARA

▶ And thanks to:

- ▶ Pieter van Beek
- ▶ Machiel Jansen
- ▶ Maurice Bouwhuis