

*Bridging the High
Performance Computing
Gap: The OurGrid
Experience*

Walfredo Cirne
walfredo@dsc.ufcg.edu.br

eScience and Grid

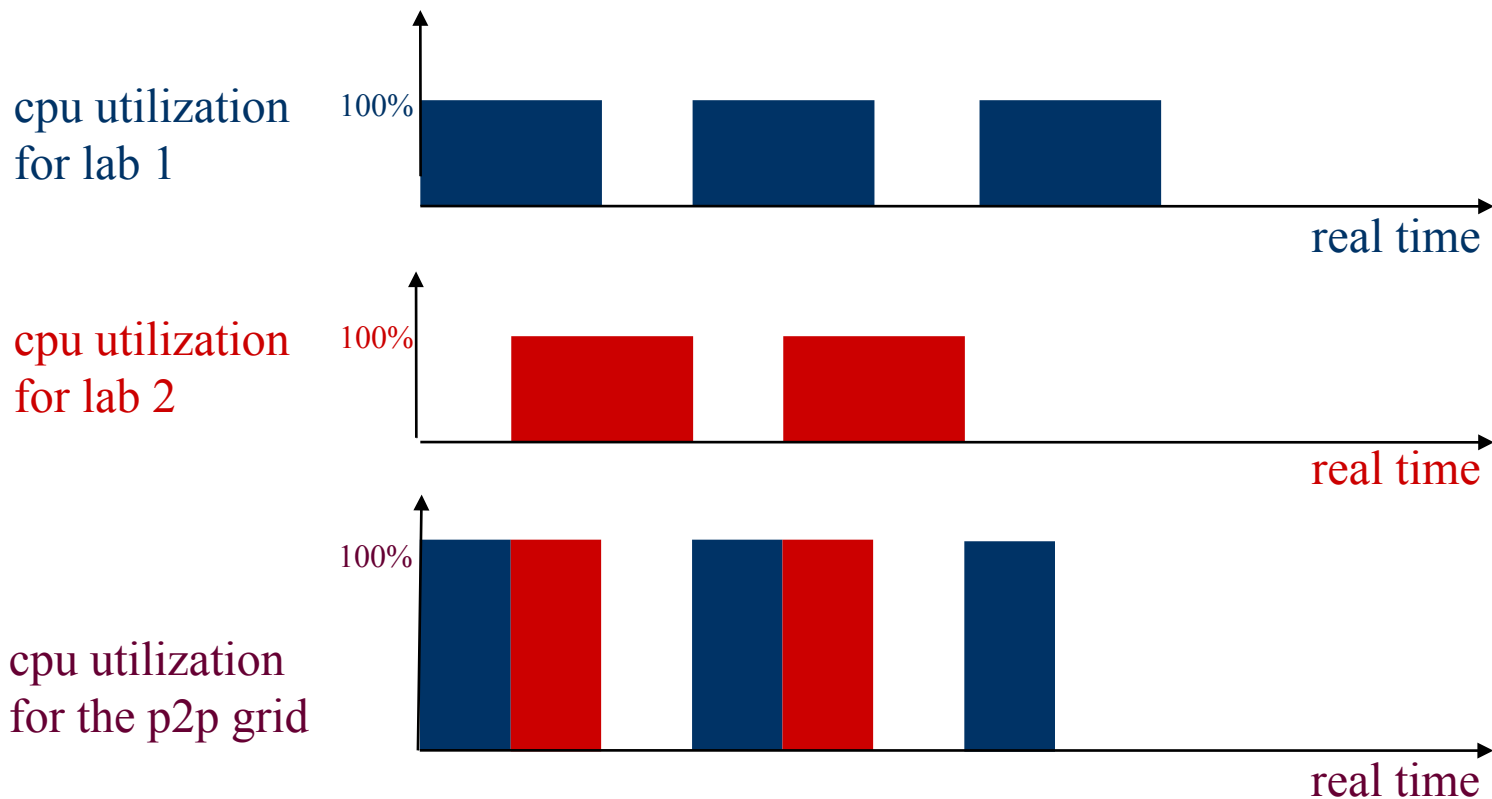
- Computers are changing scientific research
 - Enabling collaboration
 - As investigation tools (simulations, data mining, etc...)
- As a result, many research labs around the world are now computation hungry
- Grid Computing has emerged the solution to enable eScience
- Alas, it's been very hard to implement the vision of “plug and solve your problem”

And what about the thousands of small and middle research labs throughout the world which also need to be eScience-enabled?

A different trade-off: OurGrid

- OurGrid is a peer-to-peer grid
 - Each lab correspond to a peer in the system
 - Labs can freely join the system without any human intervention
 - OurGrid is easy to install and automatically configures itself
- To keep it doable, we focus on Bag-of-Tasks application

Rational of a peer-to-peer grid



Bag-of-Tasks Applications

- Data mining
- Massive search (as search for crypto keys)
- Parameter sweeps
- Monte Carlo simulations
- Fractals (such as Mandelbrot)
- Image manipulation (such as tomography)
- And many others...

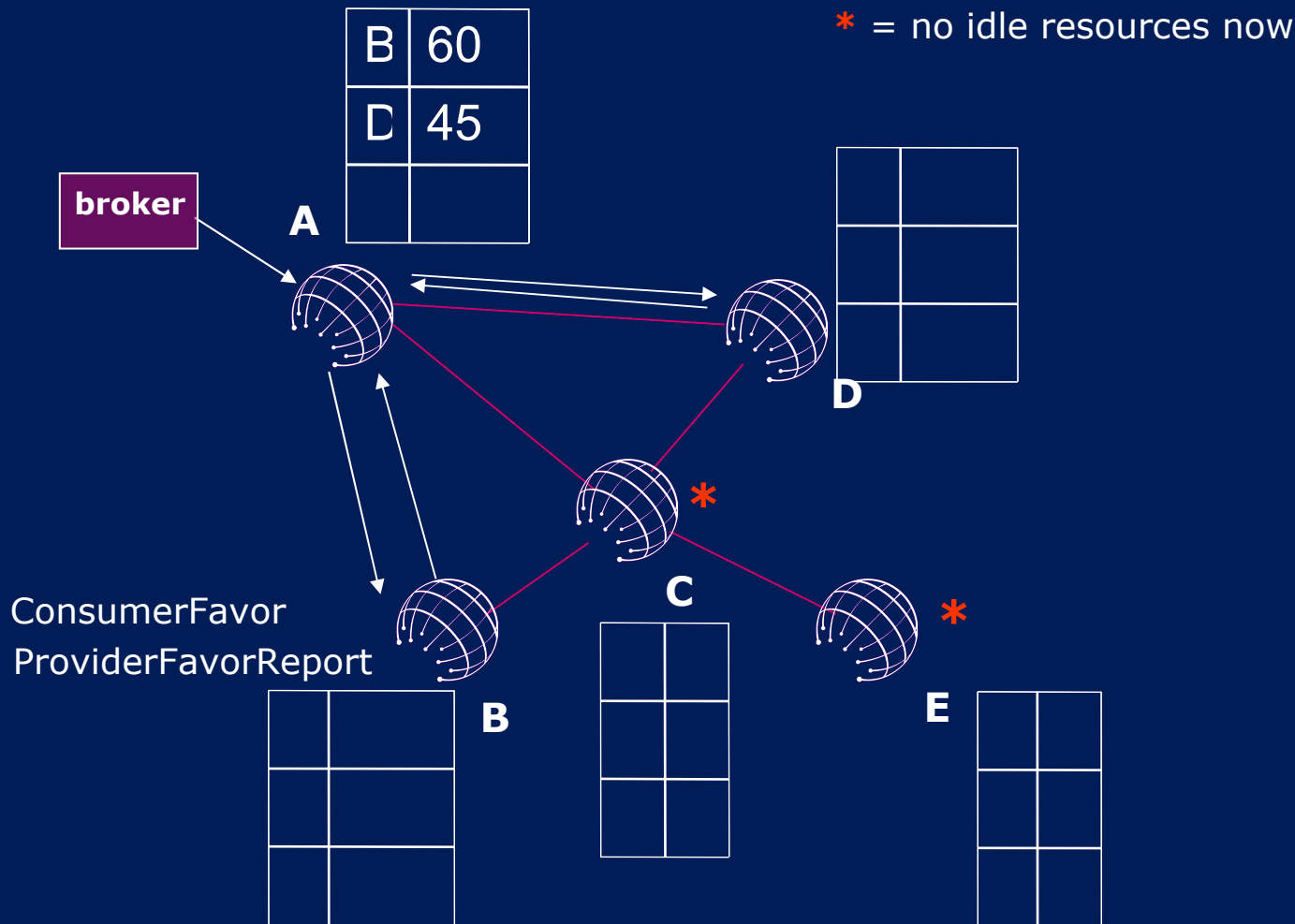
OurGrid Challenges

- How to make people collaborate?
 - Free-riders are the norm in peer-to-peer networks
 - Why should you collaborate with someone you don't know?
- How to keep it simple?
 - Grids are complex for a reason
 - How to deal with the need for information and configuration?
- How to keep it safe?
 - Labs you don't know (or trust) can freely join the grid

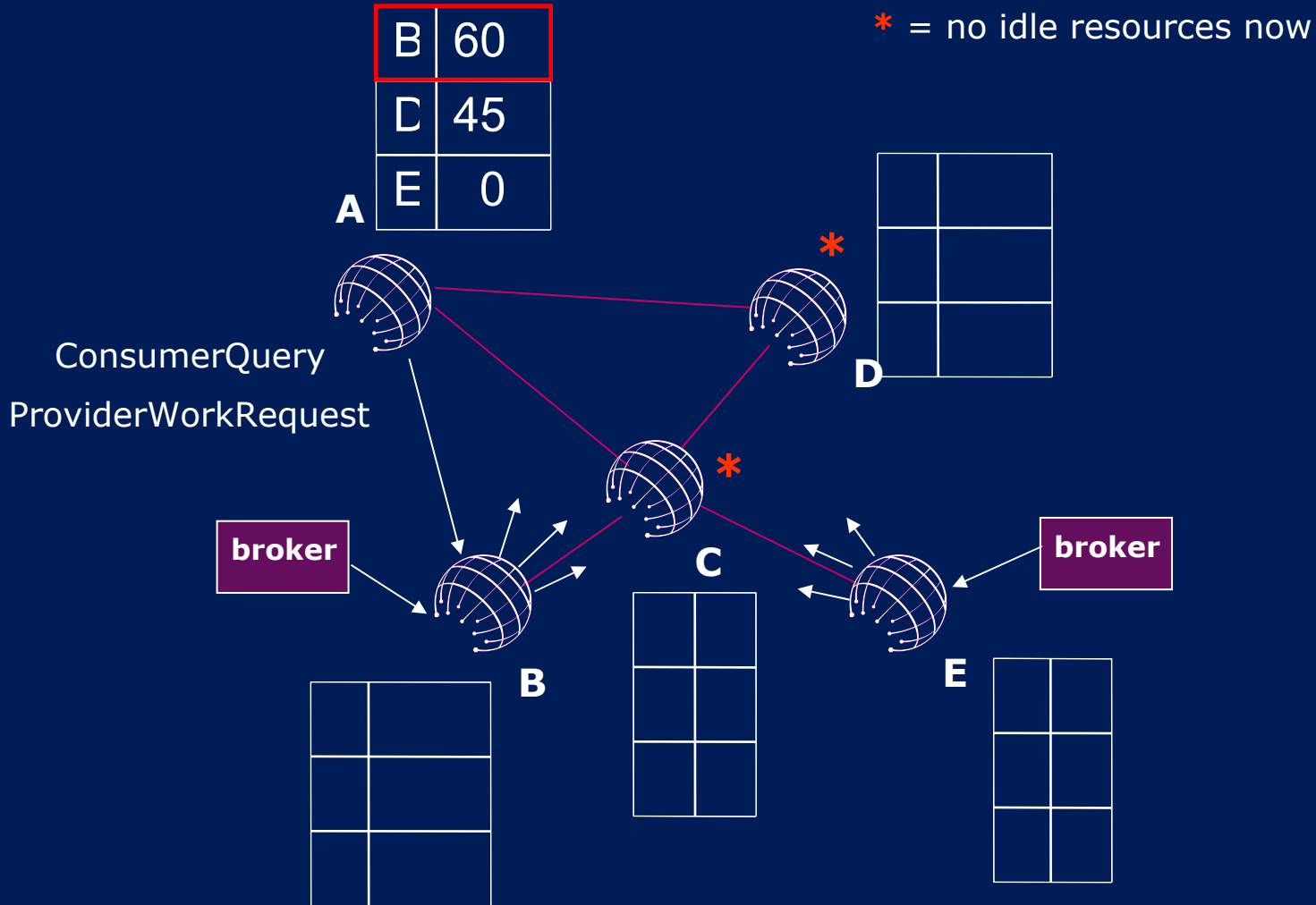
Network of Favors

- OurGrid forms a peer-to-peer community in which **peers are free to join**
- It's important to encourage collaboration within OurGrid (i.e., resource sharing)
 - In file-sharing, most users **free-ride**
- OurGrid uses the **Network of Favor**
 - All peers maintain a **local** balance for all known peers
 - Peers with greater balances have priority
 - The emergent behavior of the system is that by donating more, you get more resources
 - **No additional infrastructure is needed**

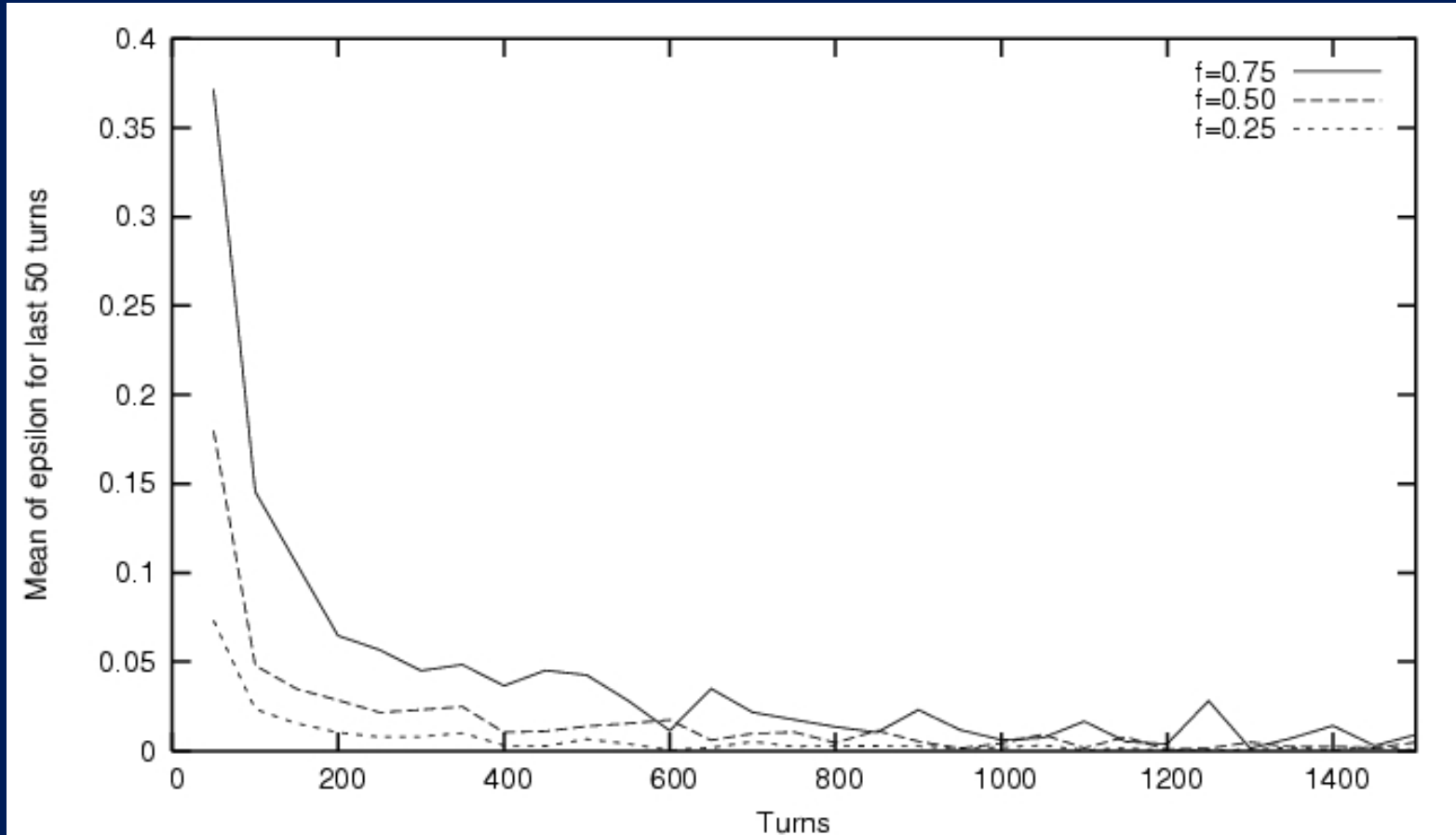
NoF at work [1]



NoF at work [2]

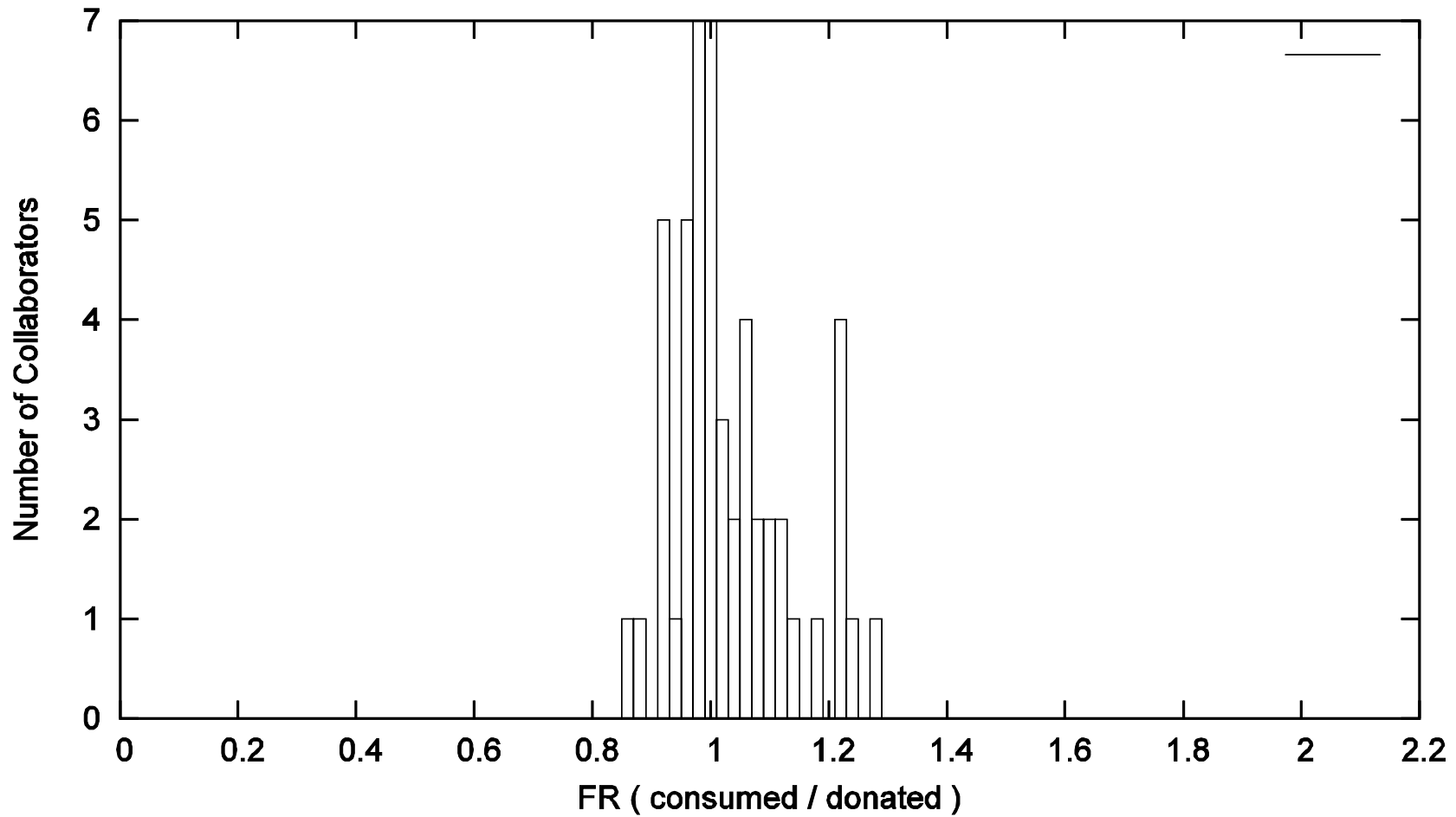


Free-rider consumption



- Epsilon is the fraction of resources consumed by free-riders

Equity among Collaborators



Conditions for NoF to work

- Moderate scale
 - Things start to break with tens of thousands peers
- Symmetric interest
 - Or, more strongly, few goods being traded
- Collaborators stay in the system for a long time
- That is why NoF does not solve the free-riding problem in file-sharing peer-to-peer system

References on NoF

- A Reciprocation-Based Economy for Multiple Services
Miranda Mowbray, Francisco Brasileiro, Nazareno Andrade, Jaudson Santana, Walfredo Cirne,
6th IEEE International Conference on Peer-to-Peer Computing (P2P'2006)
June 2006
- Accurate Autonomous Accounting in Peer-to-Peer Grids
Robson Santos, Alisson Andrade, Walfredo Cirne, Francisco Brasileiro, Nazareno Andrade
3rd Workshop on Middleware for Grid Computing (MGC2005)
November 2005
- Discouraging Free Riding in a Peer-to-Peer CPU-Sharing Grid
Nazareno Andrade, Francisco Brasileiro, Walfredo Cirne, Miranda Mowbray
13th High Performance Distributed Computing Symposium (HPDC'2004)
June 2004
- **When Can an Autonomous Reputation Scheme Discourage Free-riding in a Peer-to-Peer System?**
Nazareno Andrade, Miranda Mowbray, Walfredo Cirne, Francisco Brasileiro
4th International Workshop on Global and Peer-to-Peer Computing
April 2004
- OurGrid: An Approach to Easily Assemble Grids with Equitable Resource Sharing
Nazareno Andrade, Walfredo Cirne, Francisco Brasileiro, Paulo Roisenberg
9th Workshop on Job Scheduling Strategies for Parallel Processing
June 2003

Scheduling with No Information

- Grid scheduling typically depends on information about the grid (e.g. machine speed and load) and the application (e.g. task size)
- However, getting good information is hard
- Can we schedule without information and deploy the system now?
 - This would make the system much **easier to deploy** and **simple to use**

Work-queue with Replication

- Tasks are sent to idle processors
- When there are no more tasks, running tasks are replicated on idle processors
- The first replica to finish is the official execution
- Other replicas are cancelled

Evaluation

- 8000 experiments
- Experiments varied in
 - grid heterogeneity
 - application heterogeneity
 - application granularity
- Performance summary:

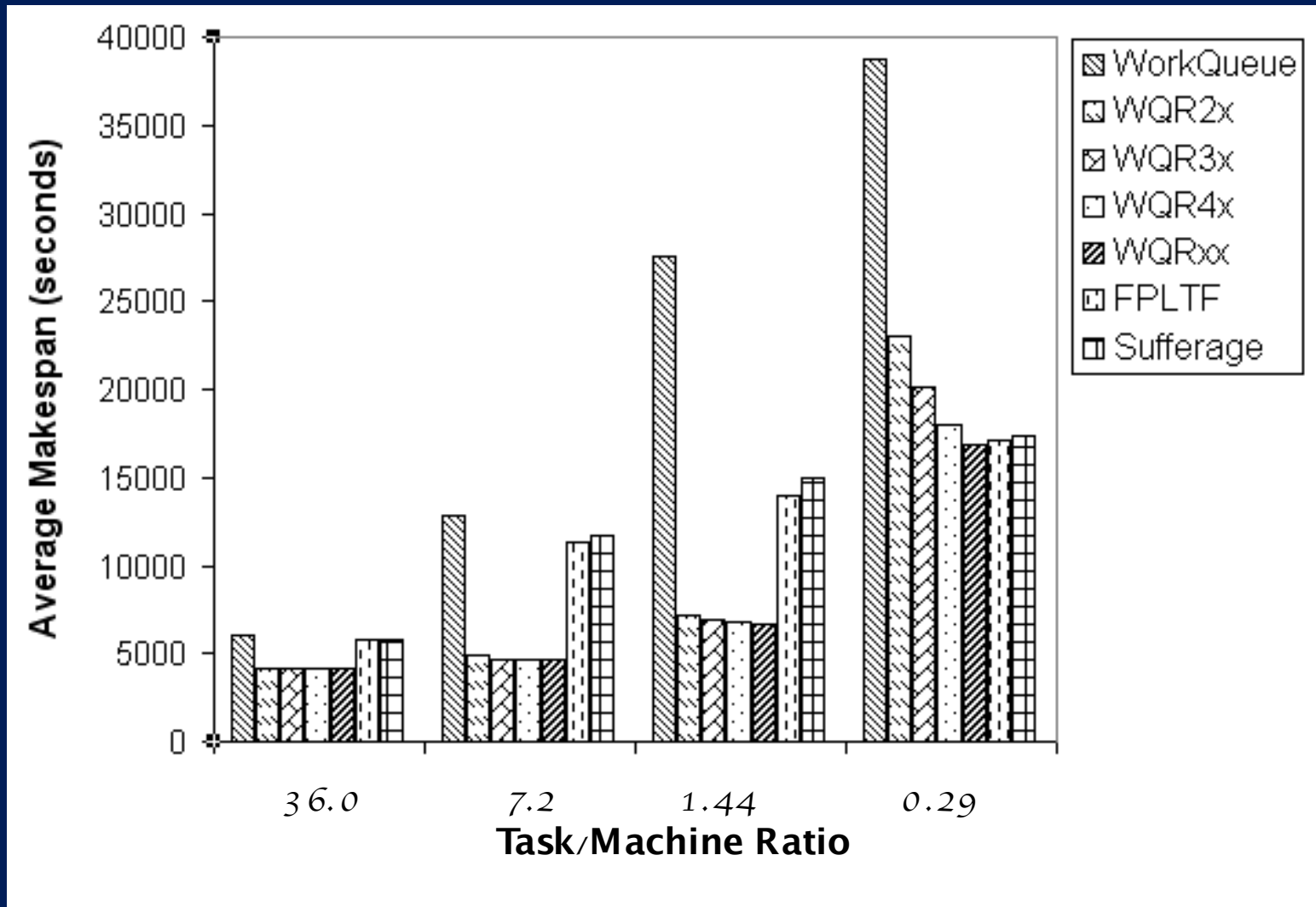
| | Sufferage | DFPLTF | Workqueue | WQR 2x | WQR 3x | WQR 4x |
|-----------|-----------|----------|-----------|----------|----------|----------|
| Average | 13530.26 | 12901.78 | 23066.99 | 12835.70 | 12123.66 | 11652.80 |
| Std. Dev. | 9556.55 | 9714.08 | 32655.85 | 10739.50 | 9434.70 | 8603.06 |

WQR Overhead

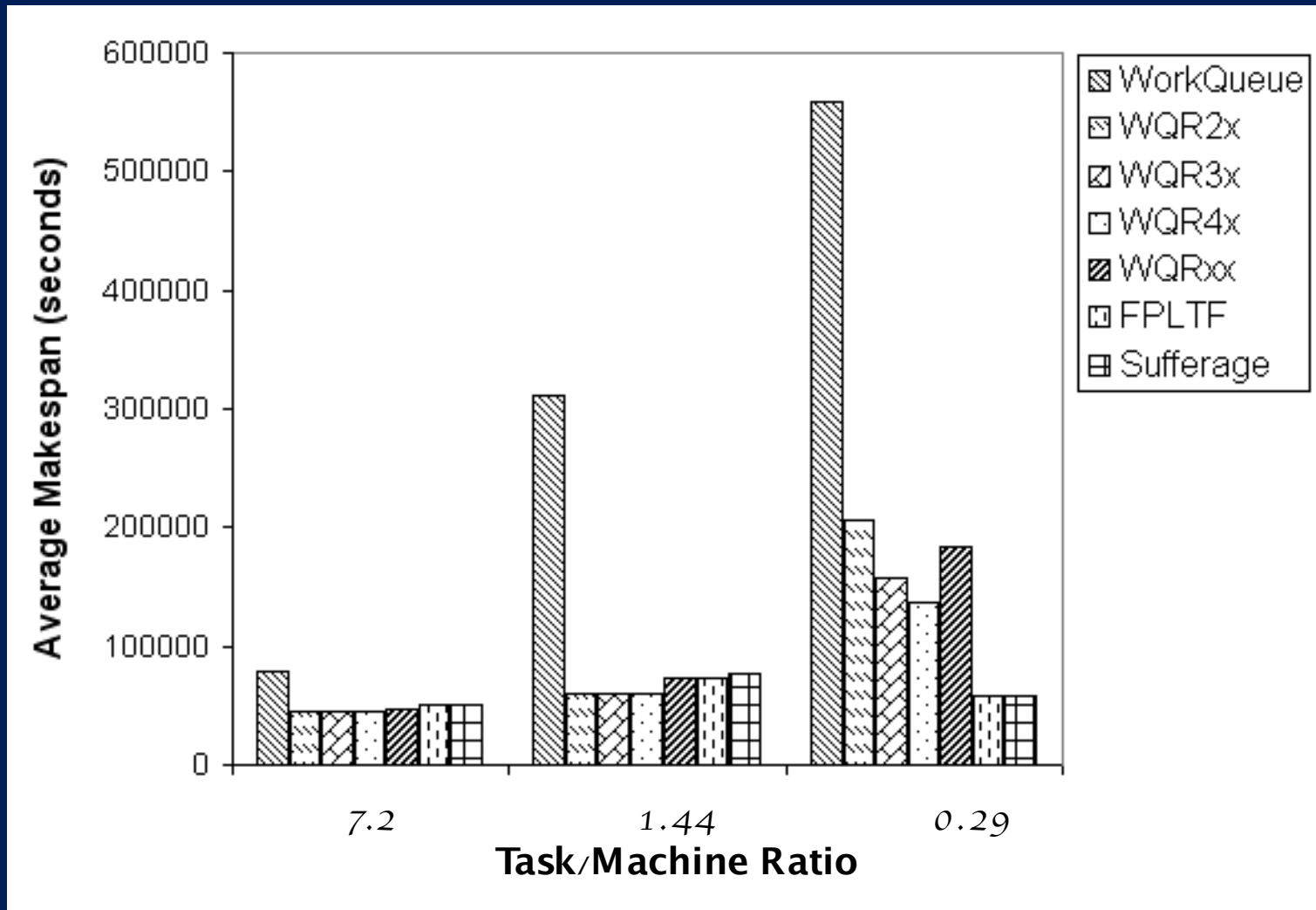
- Obviously, the drawback in WQR is cycles wasted by the cancelled replicas
- Wasted cycles:

| | WQR 2x | WQR 3x | WQR 4x |
|------------------|---------------|---------------|---------------|
| Average | 23.55% | 36.32% | 48.87% |
| Std. Dev. | 22.29% | 34.79% | 48.93% |

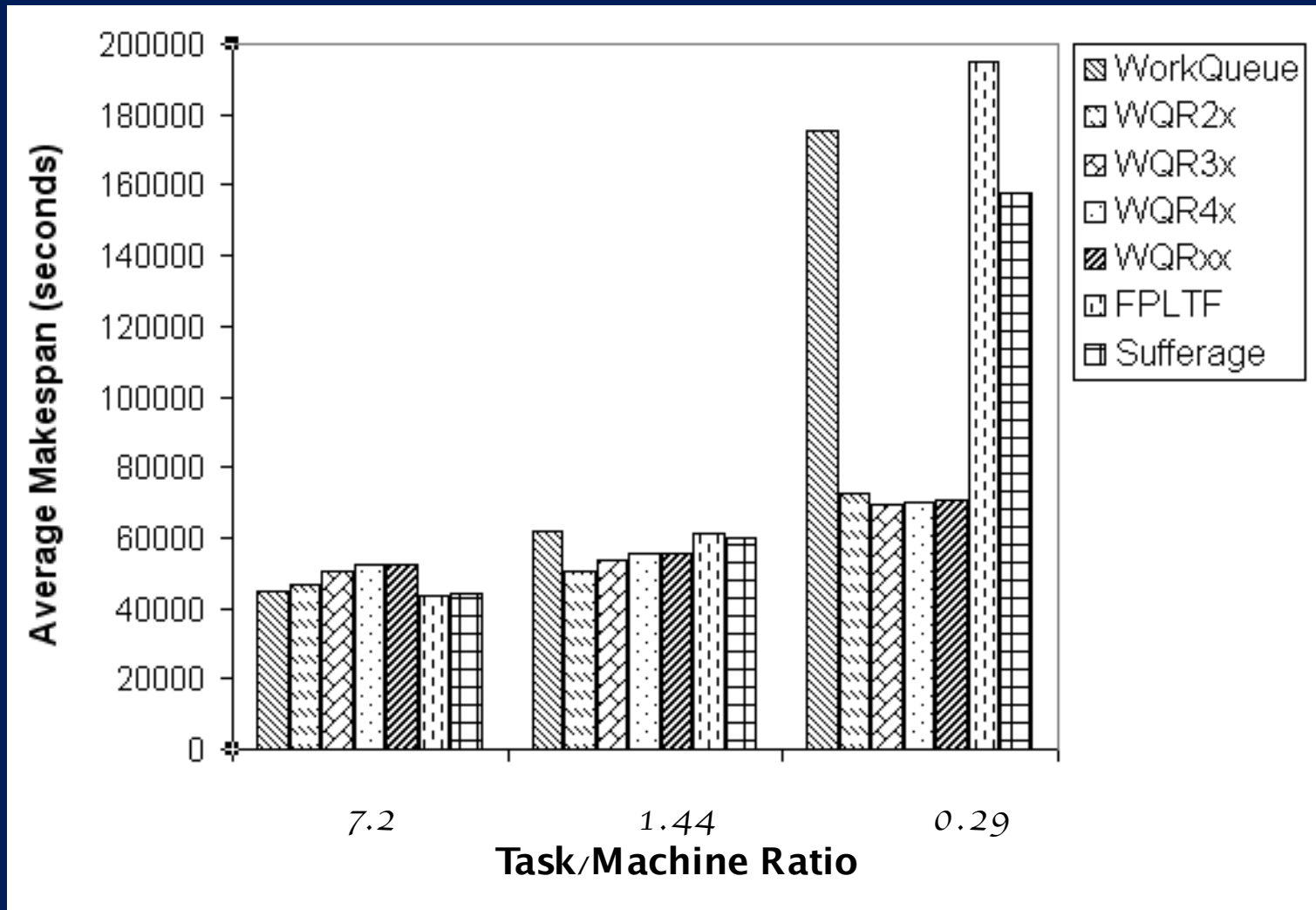
Task/Machine Performance



What if we have a free for all?



... but this can easily be solved with a **local** referee!!!



Data Aware Scheduling

- WQR achieves good performance for CPU-intensive BoT applications
- However, many important BoT applications are data-intensive
- These applications frequently reuse data
 - During the same execution
 - Between two successive executions

Storage Affinity

- Storage Affinity uses replication and just a bit of static information to achieve good scheduling for data intensive applications
- Storage Affinity uses information on which data servers have already stored a data item

Storage Affinity Results

- 3000 experiments
- Experiments varied in
 - grid heterogeneity
 - application heterogeneity
 - application granularity
- Performance summary:

| | Storage Affinity | X-Suffrage | WQR |
|--------------------|------------------|------------|---------|
| Average (seconds) | 57.046 | 59.523 | 150.270 |
| Standard Deviation | 39.605 | 30.213 | 119.200 |

References on Replication Scheduling

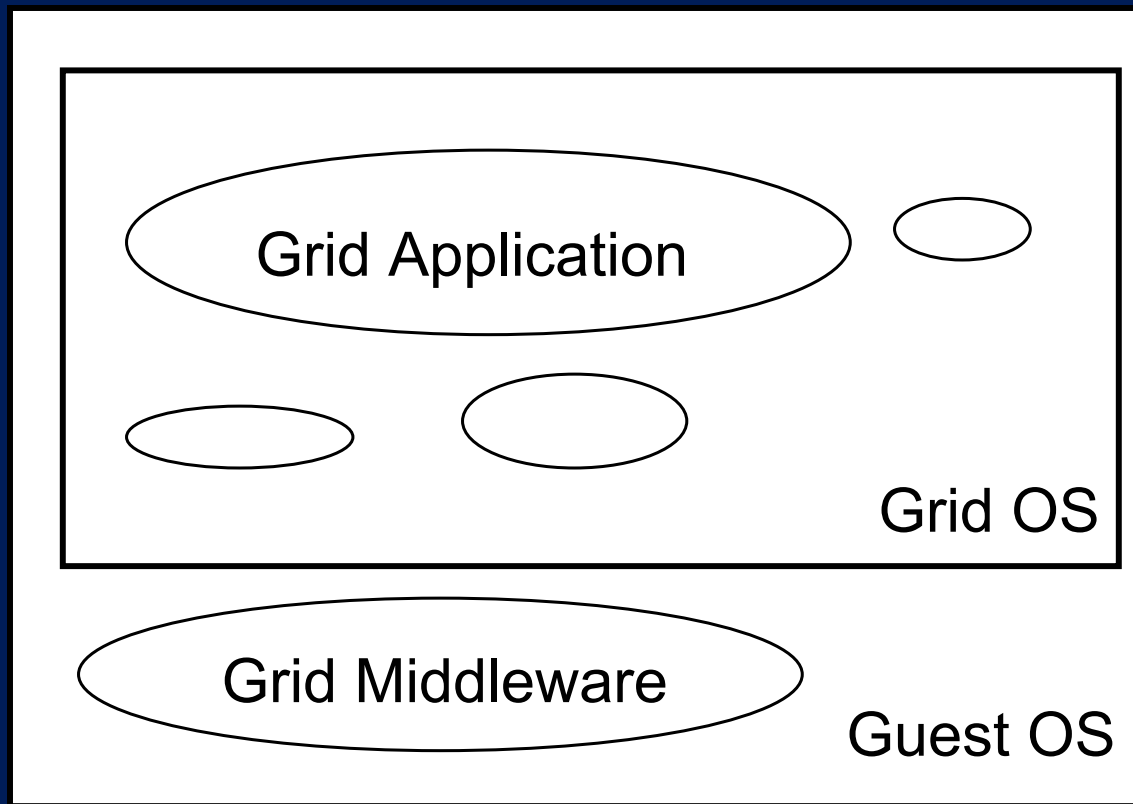


- On the Efficacy, Efficiency and Emergent Behavior of Task Replication in Large Distributed Systems
Walfredo Cirne, Daniel Paranhos, Francisco Brasileiro, Luís Fabrício W. Góes, William Voorsluys
Parallel Computing, vol. 33, no. 3, pages 213-234
April 2007
- Exploiting Replication and Data Reuse to Efficiently Schedule Data-intensive Applications on Grids
Elizeu Santos-Neto, Walfredo Cirne, Francisco Brasileiro, Aliandro Lima
10th Workshop on Job Scheduling Strategies for Parallel Processing
June 2004
- Trading Cycles for Information: Using Replication to Schedule Bag-of-Tasks Applications on Computational Grids
Daniel Paranhos, Walfredo Cirne, Francisco Brasileiro
Euro-Par 2003: International Conference on Parallel and Distributed Computing
August 2003

SWAN: OurGrid Security

- Running an unknown application that comes from an unknown peer is a clear security threat
- We leverage the fact that Bag-of-Tasks applications can (and typically are) written communicate via input/output files
- **Input/output is done by OurGrid itself**
- The remote task runs inside a Xen virtual machine, with no network access, and disk access only to a designated partition

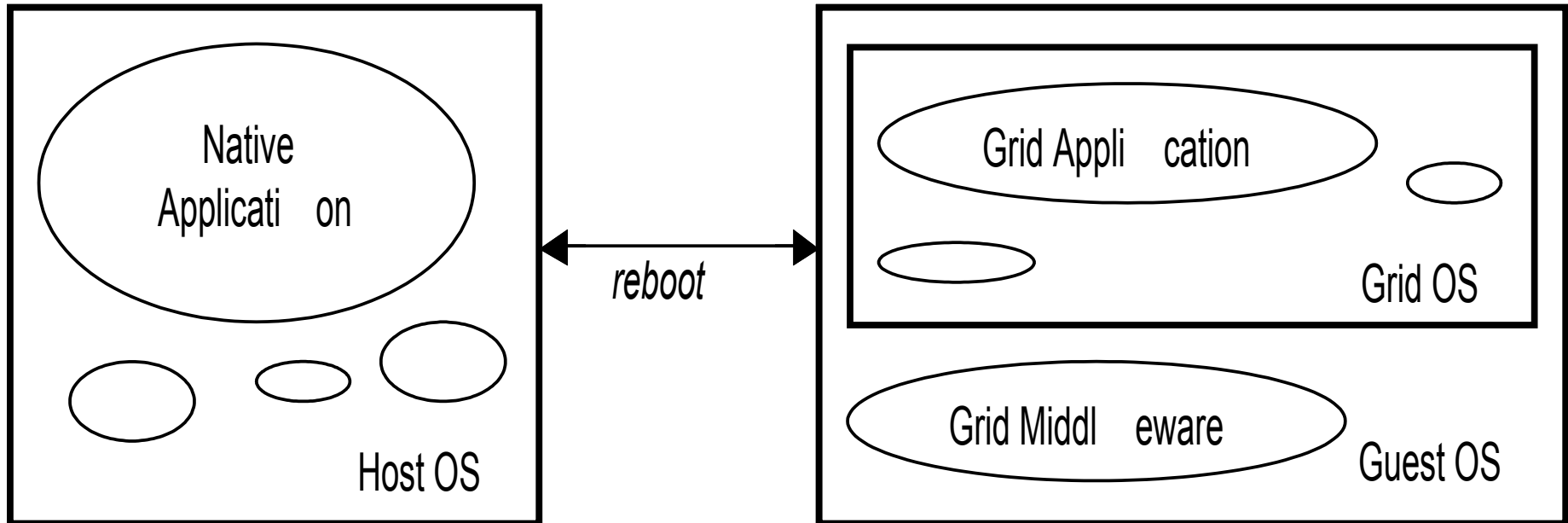
SWAN Architecture



Adding a second line of defense

- We can also reboot to add a second layer of protection to the user data and resources
- This has the extra advantage of enabling us to use an OS different from that chosen by the user
 - That is, even if the interactive user prefers Windows, we can still have Linux
- Booting back to the user OS can be done fast by using hibernation

SWAN + Mode Switcher



One size does not fit all

- Currently extending SWAN to support other virtualization technologies
 - vserver, VirtualBox, VMWare

Attacks to the application

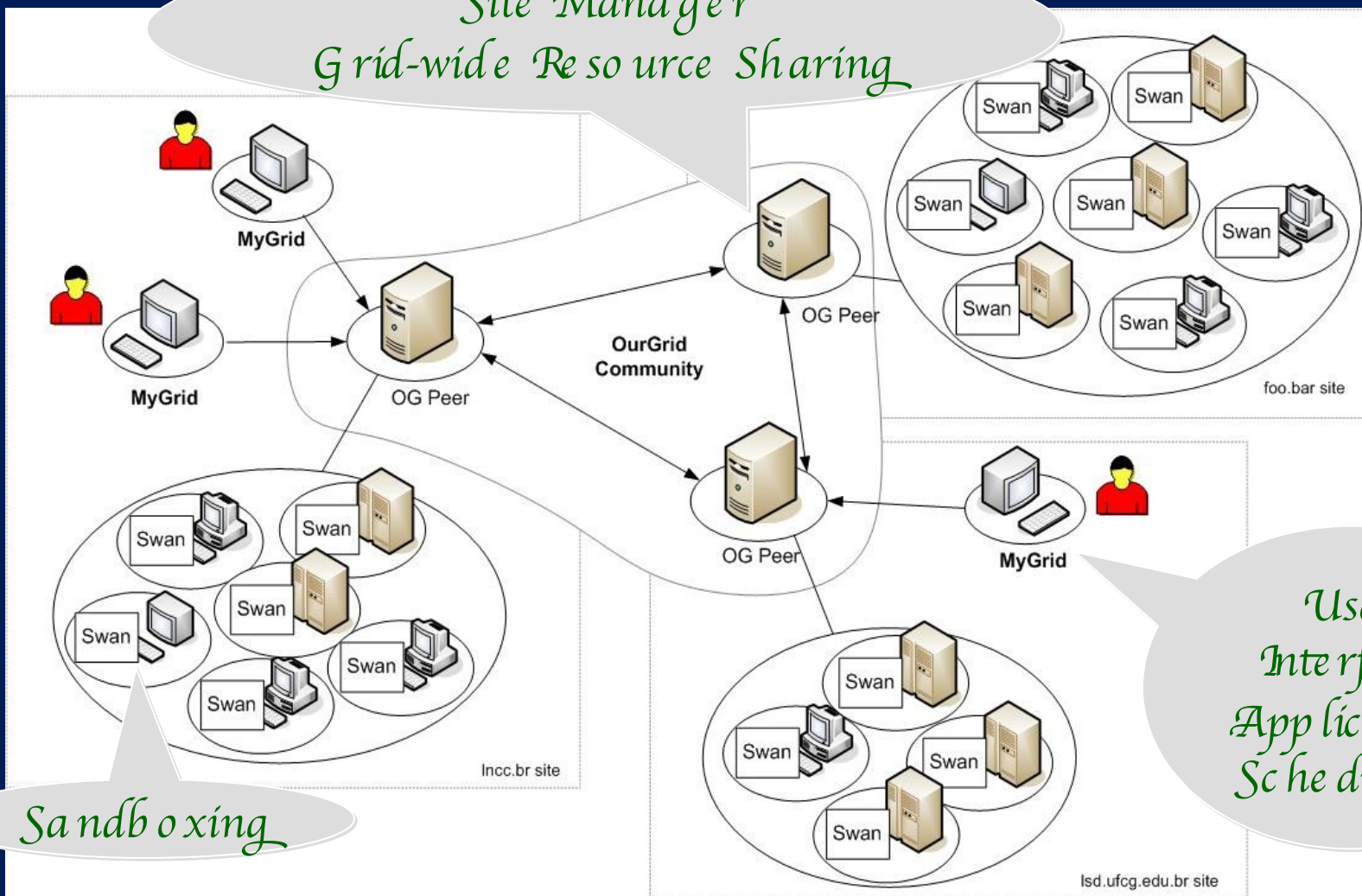
- Application may introduce task-dependent water marks
- Credibility-based sabotage tolerance
 - Generic, but more expensive
 - Based on the work of Luis Samernta

SWAN Reference

- Sandboxing for a Free-to-join Grid with Support for Secure Site-wide Storage Area
Edjozane Cavalcanti, Leonardo Assis, Matheus Gaudêncio, Walfredo Cirne, Francisco Brasileiro, Reynaldo Novaes
First International Workshop on Virtualization Technology in Distributed Computing (VTDC 2006)
November 2006

OurGrid Architecture

*Site Manager
Grid-wide Resource Sharing*



Sandboxing

*User
Interface
Application
Scheduling*

Other “details”

- A very simple scripting language can be used to describe tasks
 - Based on put/get files to/from the remote machine
 - A Java API can also be used
 - Attribute constraints and preferences are available
- Firewalls/NATs are dealt with by using XMPP (Jabber) as communication infrastructure

Grid-enabling an Application

- Write a script using a very simple language
 - Simple abstractions
 - File transfer (put, store, get)
 - Hide heterogeneity (\$PLAYPEN, \$STORAGE)
 - Define constraints (job requirements and grid machine attributes)
- Write a program that embeds the business logic and may make use of more complex features available through a Java API
- Deploy a Portal that embeds the application

An Example: Factoring with OurGrid

job:

label: my_factorial_useless_example
requirements: (os=linux)

task:

init: store ./Fat.class \$PLAYPEN
remote: java Fat 3 18655 34789789799 output-\$JOB-\$TASK
final: get \$PLAYPEN/output-\$JOB-\$TASK results

task:

init: store ./Fat.class \$PLAYPEN
remote: java Fat 18656 37307 34789789799 output-\$JOB-\$TASK
final: get \$PLAYPEN/output-\$JOB-\$TASK results

task:

init: store ./Fat.class \$PLAYPEN
remote: java Fat 37308 55968 34789789799 output-\$JOB-\$TASK
final: get \$PLAYPEN/output-\$JOB-\$TASK results

...

OurGrid Status

- OurGrid supports the OurGrid Community
 - A free-to-join grid that is in production since December 2004 [see status.ourgrid.org]
- It also supports ShareGrid
 - A collaborative project, coordinated by TOPIX (TOrino Piemonte Internet eXchange) in Italy [see dcs.di.unipmn.it]
- OurGrid is **open source** (GPL)
 - Version 4.0 has just been released!
 - Contributions are welcome!



OurGrid

ourgrid.org visits



- More than 10,000 downloads since December 2004


http://status.ourgrid.org

OurGrid Web Status 3.2.1 - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://status.ourgrid.org/

Globo Online - Morales diz ... The Online Photographer Digital Camera Reviews an... OurGrid Web Status 3... Chamada_Publica_MCT_F...



OurGrid
STATUS

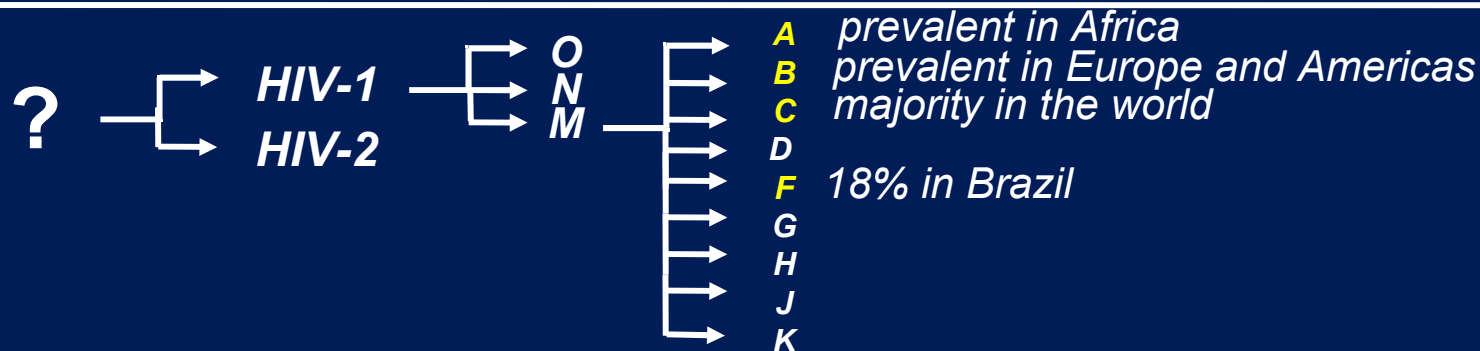
Statistics

Server time: **Tue May 09 10:44:44 BRT 2006**
 Last snapshot time: **Tue May 09 10:43:24 BRT 2006**
 Peers: **13**
 Grid machines: **310**
 Online Grid machines: **202**

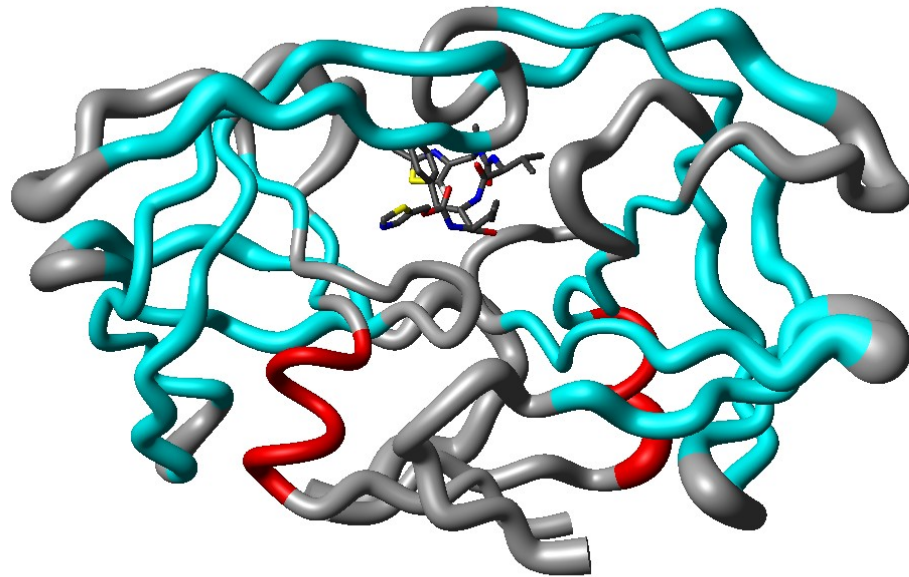
Online Peers Trying to contact peer: [Ourgrid Experimental do CIASC](#)

| | Local Machines | | | Using | | Donating |
|--|----------------|--------|------|-------|-----------|----------|
| | Total | Online | Idle | Local | Community | |
| Embrapa Informatica Agropecuaria | 4 | 2 | 2 | 0 | 0 | 0 |
| carcara.Incc.br | 40 | 31 | 28 | 0 | 0 | 3 |
| copad-dca.dca.ufcg.edu.br | 18 | 18 | 18 | 0 | 0 | 0 |
| copad-lmrs.lmrs-semarh.ufcg.edu.br | 18 | 18 | 18 | 0 | 0 | 0 |
| copad.lsd.ufcg.edu.br | 11 | 5 | 5 | 0 | 0 | 0 |
| cpad.pucrs.br | 58 | 27 | 25 | 0 | 0 | 2 |
| eegpeer.lsd.ufcg.edu.br | 10 | 10 | 10 | 0 | 0 | 0 |
| lcc.ufcg.edu.br | 39 | 25 | 25 | 0 | 0 | 0 |
| peer.lsd.ufcg.edu.br | 46 | 18 | 18 | 0 | 0 | 0 |

HIV research with OurGrid

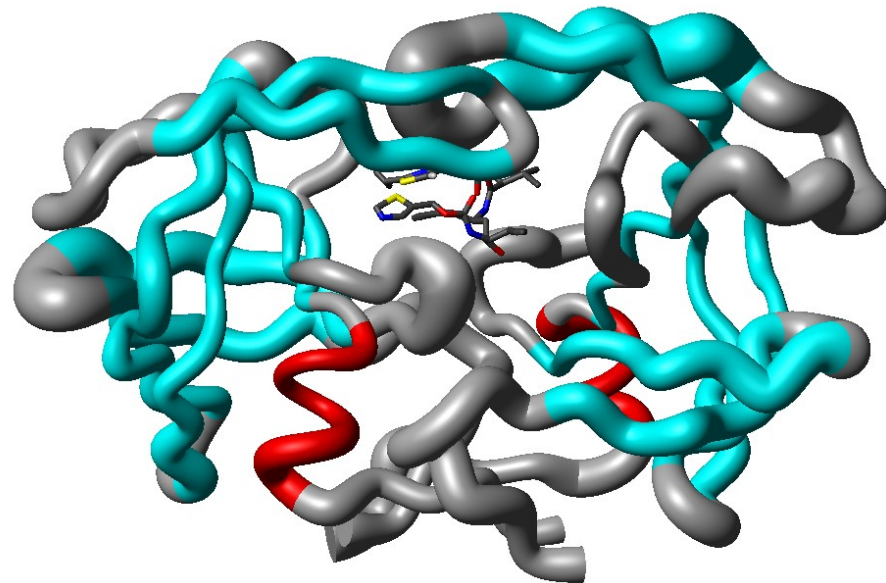


HIV protease + Ritonavir



Subtype B

Subtype F



Things that need improvement...

- The initial effort to benefit from OurGrid is greater than what we'd liked
 - Running on an unknown machine is really hard
 - Installing Xen is too intrusive
- The software should be more stable
 - Hopefully fixed with this new release
- People try to use the system for what is not designed (i.e. for non BoT application)
 - We need to think of a way to accommodate richer programming models

Conclusions

- We have an **free-to-join** grid solution for Bag-of-Tasks applications working **today**
- We have managed to combine research with solving real problems!!!
- Delivering results to real users is really cool... but a lot tougher than writing papers! :-)

If you want to read a single paper...

- Labs of the World, Unite!!!
Walfredo Cirne, Francisco Brasileiro, Nazareno Andrade,
Lauro Costa, Alisson Andrade, Reynaldo Novaes, Miranda
Mowbray
Journal of Grid Computing, vol. 4, no. 3, pages 225-246
September 2006

All papers are at...

- <http://walfredo.dsc.ufcg.edu.br/resume.html>
- <http://walfredo.cirne.net/resume.html>



OurGrid

Questions?



OurGrid

Thank you!

Merci!

Danke!

Grazie!

Gracias!

Obrigado!

谢谢!

More at www.ourgrid.org