# *Tenfold Acceleration of Office/Database/Web/Media Applications over the WAN/Grid*

**Dr. Frank Z Wang**

**Professor and Chair in Grid Computing and e-Science**
**Director, Centre for Grid Computing**
**Cambridge-Cranfield HPCF**
**http://www.hpcf.cam.ac.uk/research.html**

In partnership with IBM, Xerox, CERN, HP, Rolls Royce, EuroElectronics, and Excelian.

# Abstract:


- Our group has developed a data communication protocol, which beats classic ones tenfold over a real-world link. This protocol is the first of its kind worldwide. Best of all, this protocol requires no changes in the way users work with their applications since it conforms the existing IT infrastructures. During the presentation, videos will be presented, showing how this protocol accelerates applications, ranging from Office, Database, Web Browser to Media Player.

# Hints…

- We were inspired by GridFTP that uses multi-streams and GSI
- We have been attempting to implement a network/grid filesystem in a similar way
- Network file system is not a (disk) filesystem
- It is a data communication protocol, a platform, an infrastructure, an engine, an accelerator, …
- Other (distributed) applications can be deployed on it, like db, vi, firefox, mplayer, Google Earth…
- mount, ssi

- The success of GridFTP

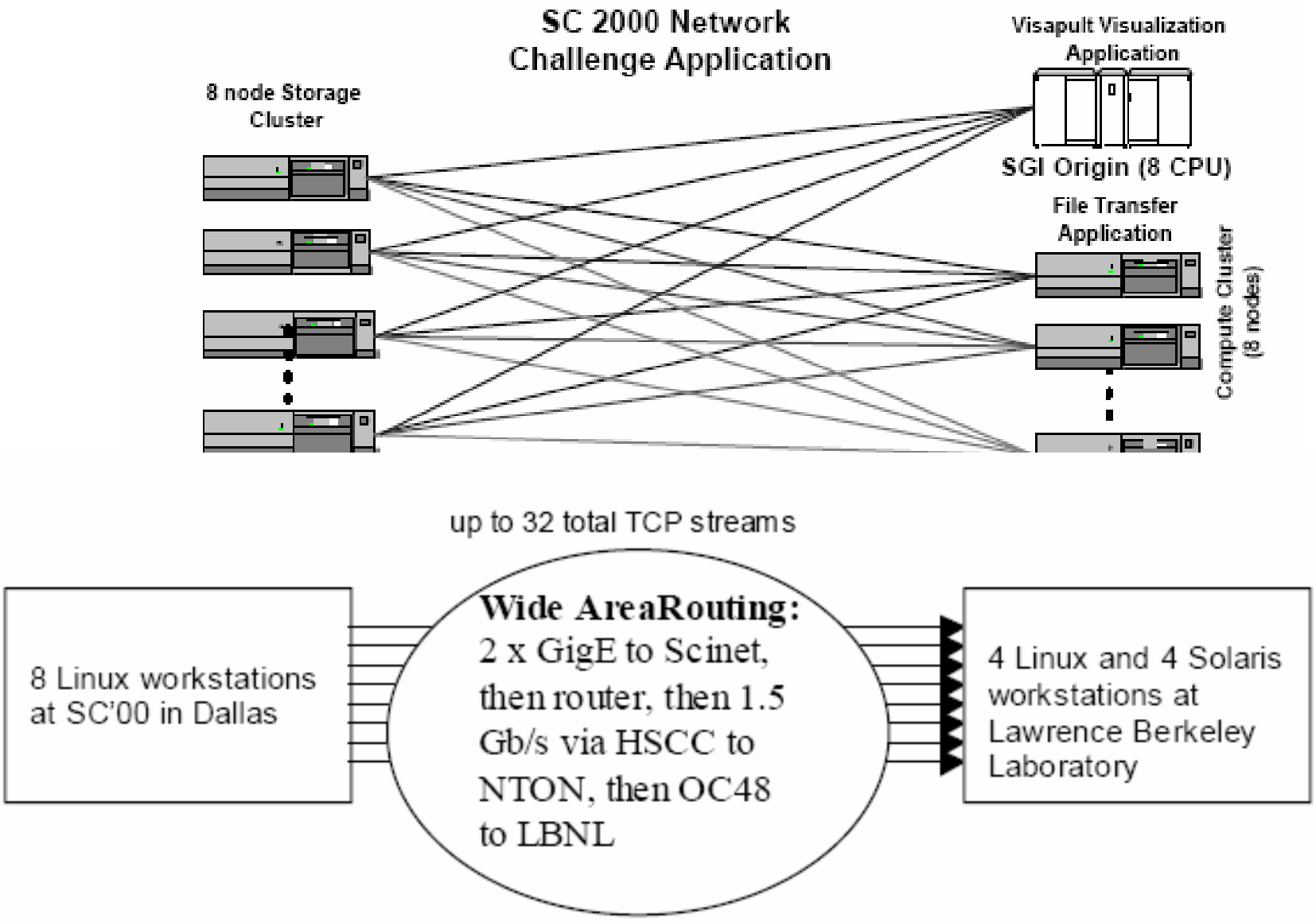
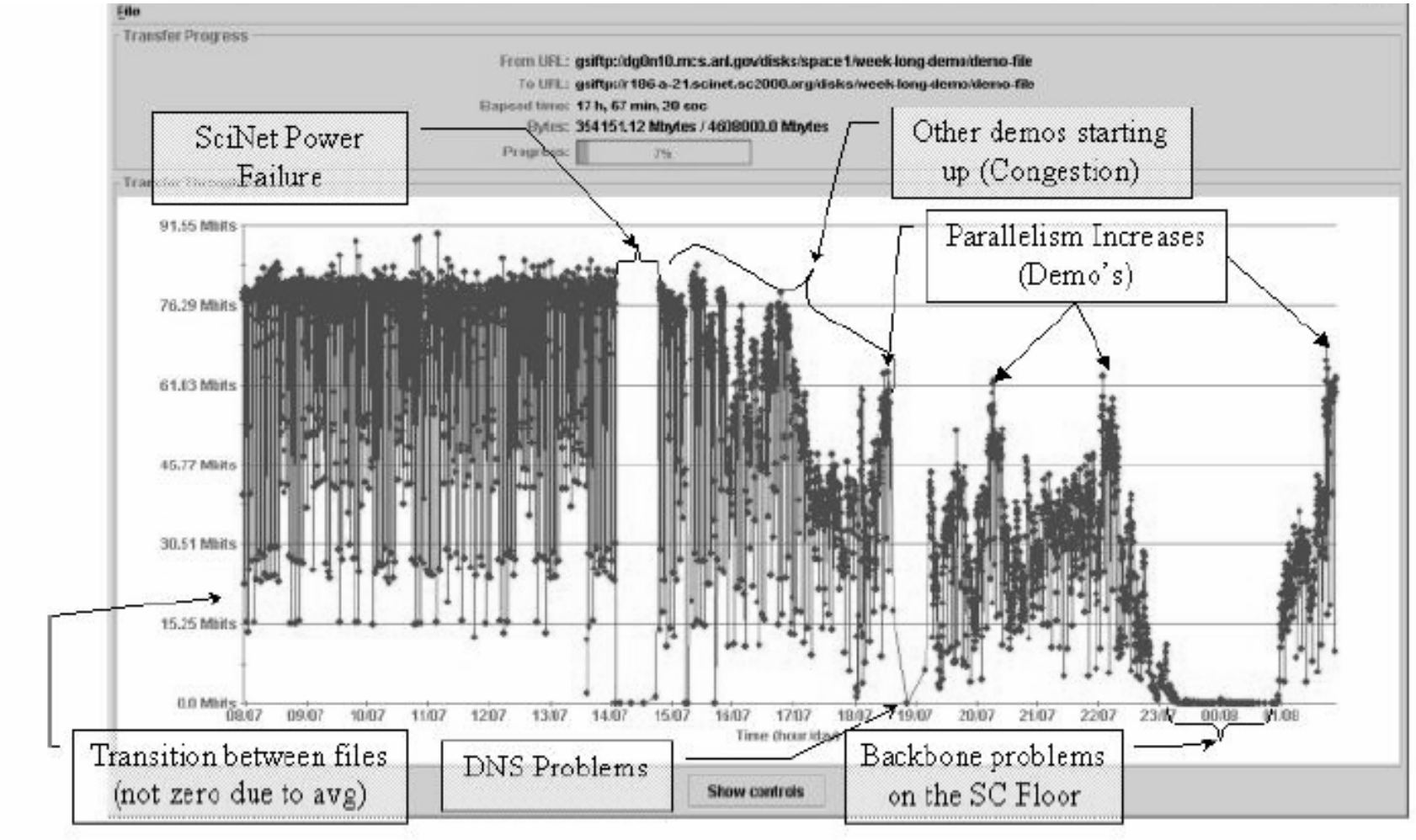



Figure 4: Experimental configuration for Network Challenge competition at SC'00.

At SuperComputing Conference (SC'00) in November 2000, achievable peak performance was measured during the Network Challenge competition over a 2 x 1Gbps link between Dallas, Texas, and Berkeley, California. A peak data rate of 1.55 gigabits/second and an average data rate of 512.9 megabits per second have been observed [].

[] Bill Allcock, Joe Bester, John Bresnahan, Ann L. Chervenak, Ian Foster, Carl Kesselman, Sam Meder, Veronika Nefedova1, Darcy Quesne, Steven Tuecke, Secure, Efficient Data Transport and Replica Management for High-Performance Data-Intensive Computing, 2000




Bandwidth measured for a series of transfers performed over a 14 hour period, between Dallas and Chicago.

# NorduGrid connectivity


NorduGrid Project
www.quark.lu.se/grid

## Standard FTP and GridFTP protocols for international data transfer in Pamela Satellite Space Experiment

R. Esposito, P. Mastroserio, G. Tortone
*INFN, Napoli, I-80126, Italy*
F. M. Taurino
*INFM, Napoli, I-80126, Italy*

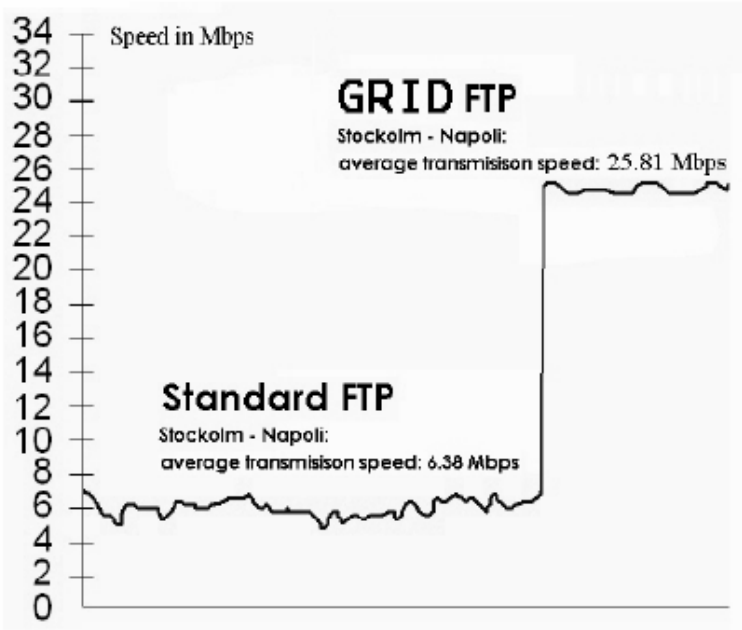



Figure 4: The average speed transferring files from Stockholm to Napoli has been of 25.81 Mbps using GridFTP.

The multithreaded GridFTP transfers resulted in a remarkable performance increase of about 600-800 percent, compared to conventional FTP downloads.

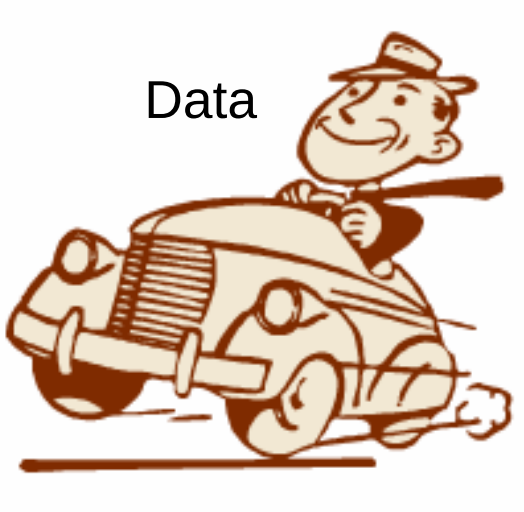
# Conclusion

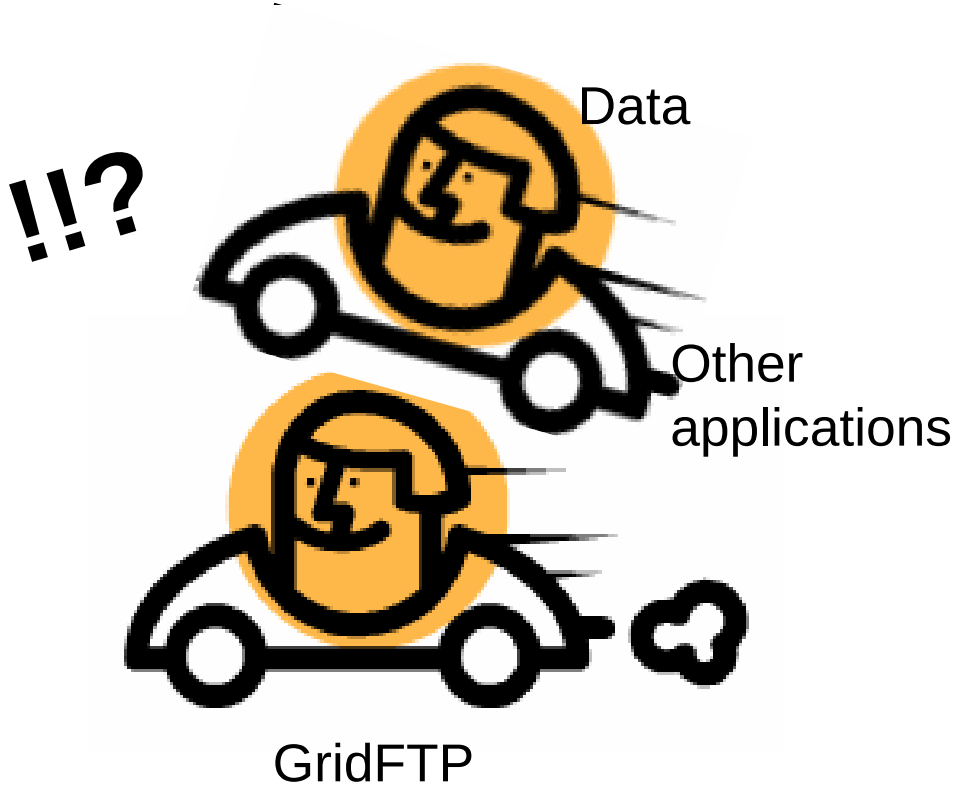
- Parallel threads radically increase performance
- ~ 20-30 threads found to be optimal
- LAN performance can be achieved over WAN
- Bottleneck: “the last mile” in your download
  - LAN NIC, client hardware (disks), machine load
- The software is stable and reliable up to ~ 60 threads
- GridFTP with parallel threads is found to be the best on the market

NorduGrid Project
www.quark.lu.se/grid

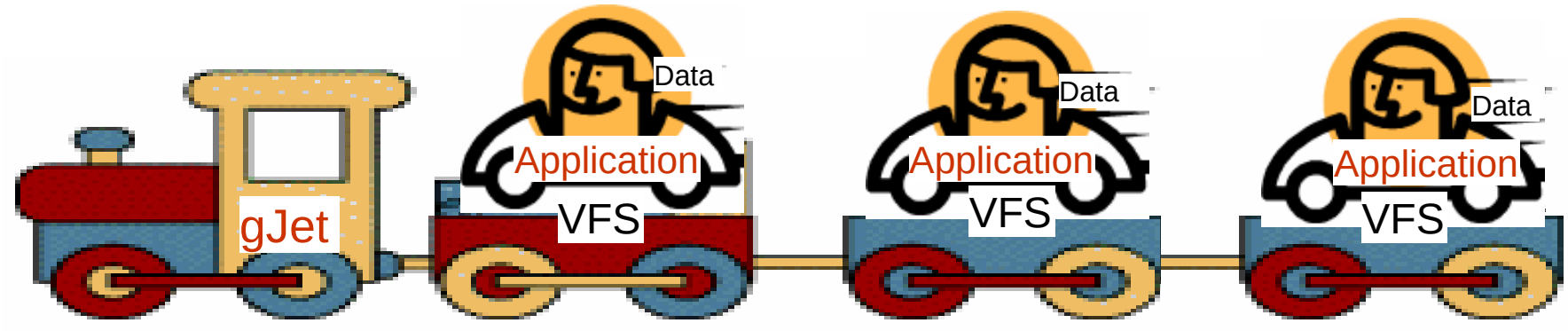
[] GridFTP tests over the NorduGrid Resources, **2nd NorduGridWorkshop** Oslo, November 1-2, 2001 Balázs Kónya

- GridFTP is faster than FTP
- But it is an independent tool to move file from one location to another
- It cannot be integrated into other applications, no API
- It cannot be used as an underlying engine to speed up other applications



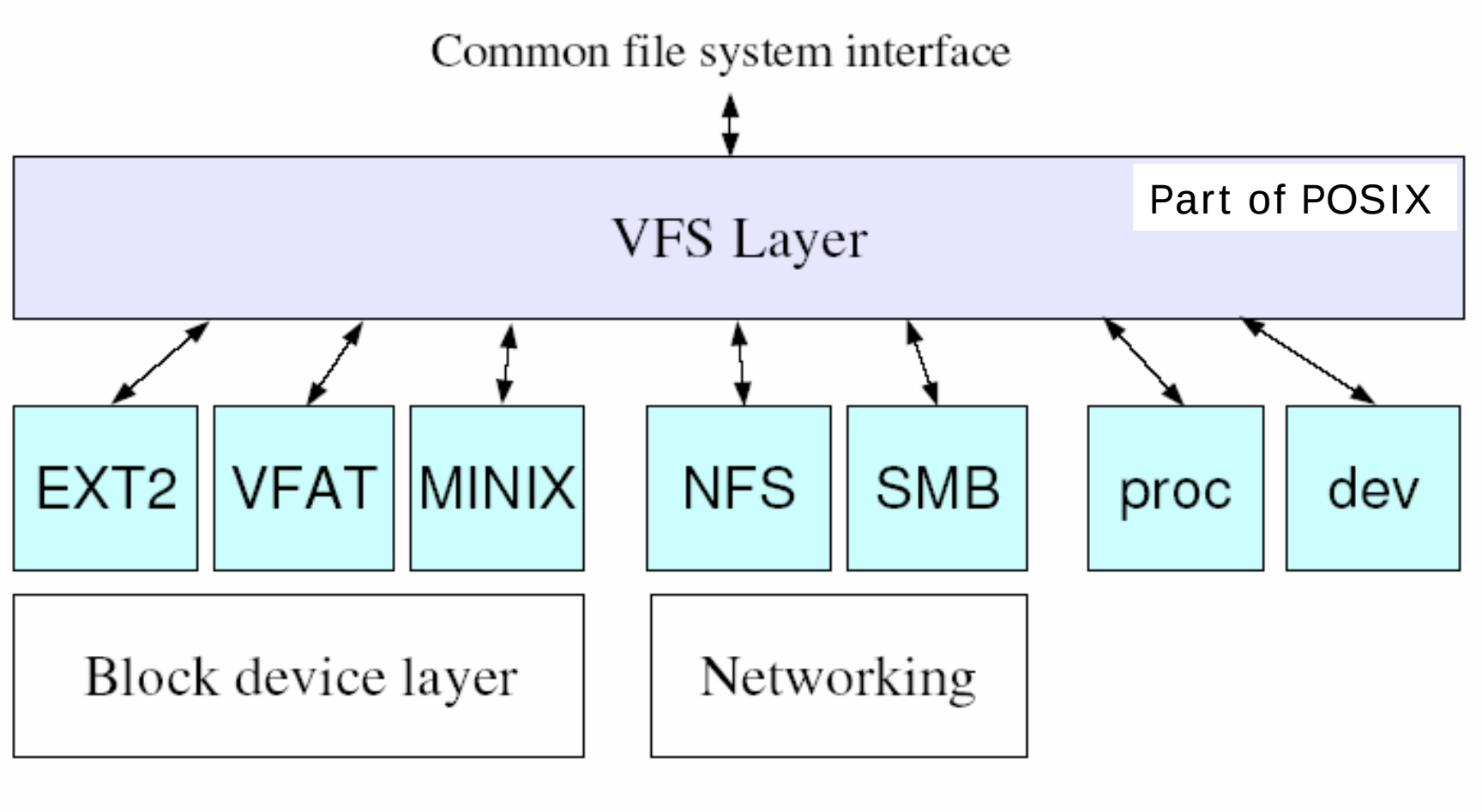



- We wanted to extend GridFTP
- gJet can be used as an underlying engine to speed up other applications



- Network filesystem differs from local (disk) filesystem

# Linux Filesystem Architecture



# How to manipulate a remote file?

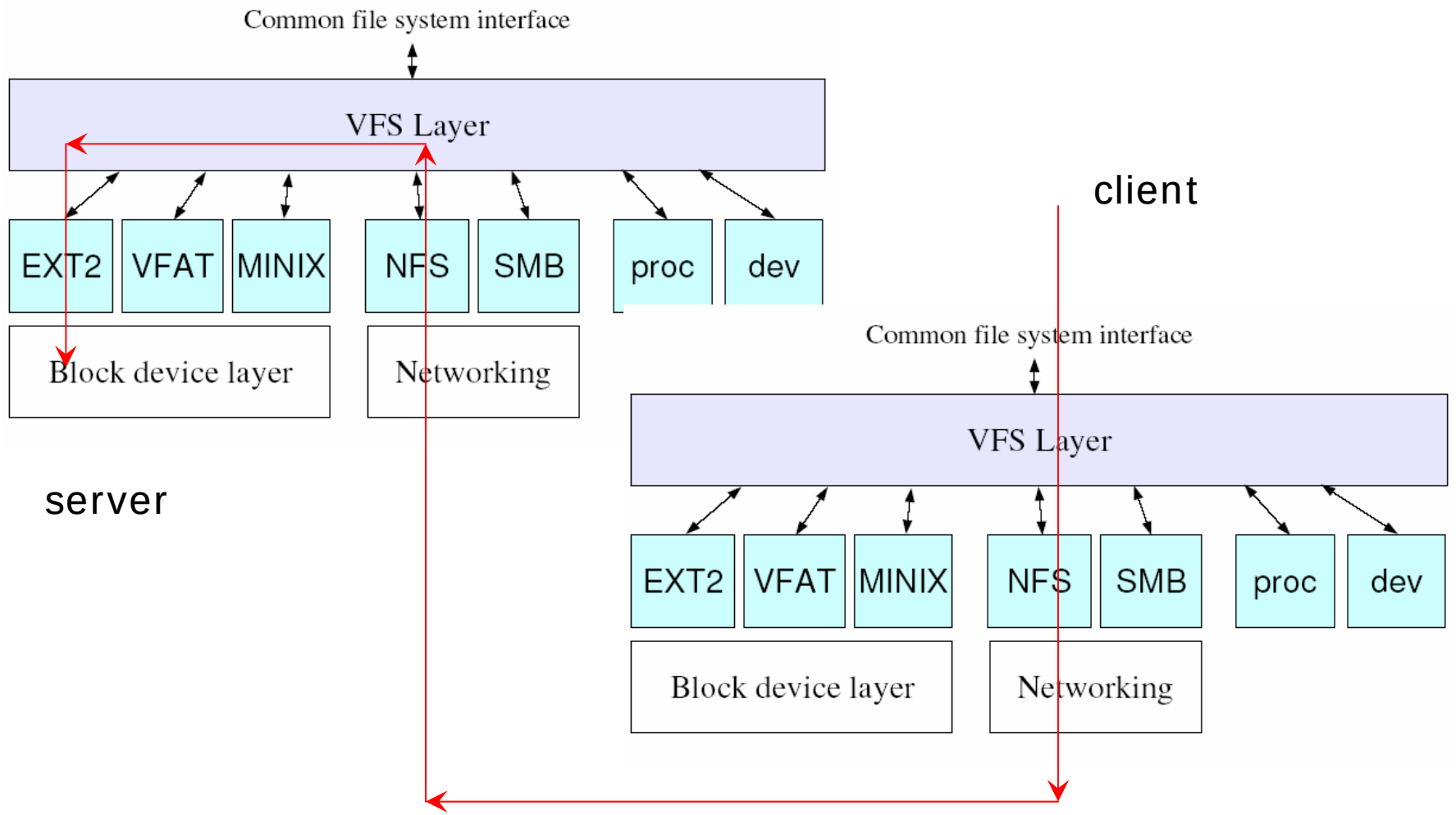


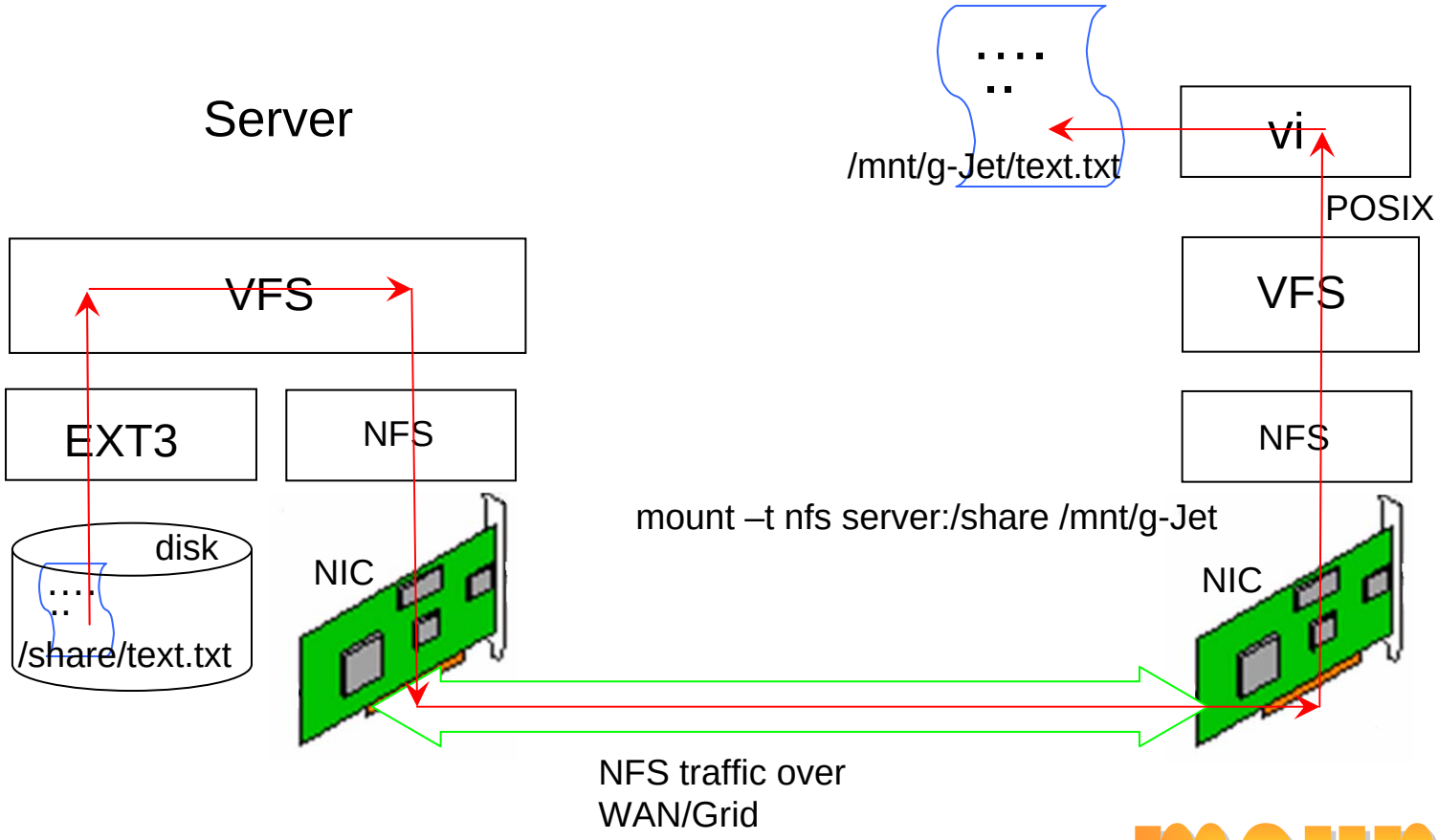

Fig.2 Different from (disk) filesystems, a "network file system" acts as a data communication protocol, providing a client with accesses to files on a remote server. Once a client mounts a file directory on a remote file server (to its local file tree) via the NFS protocol, the client can access it as if it was local.

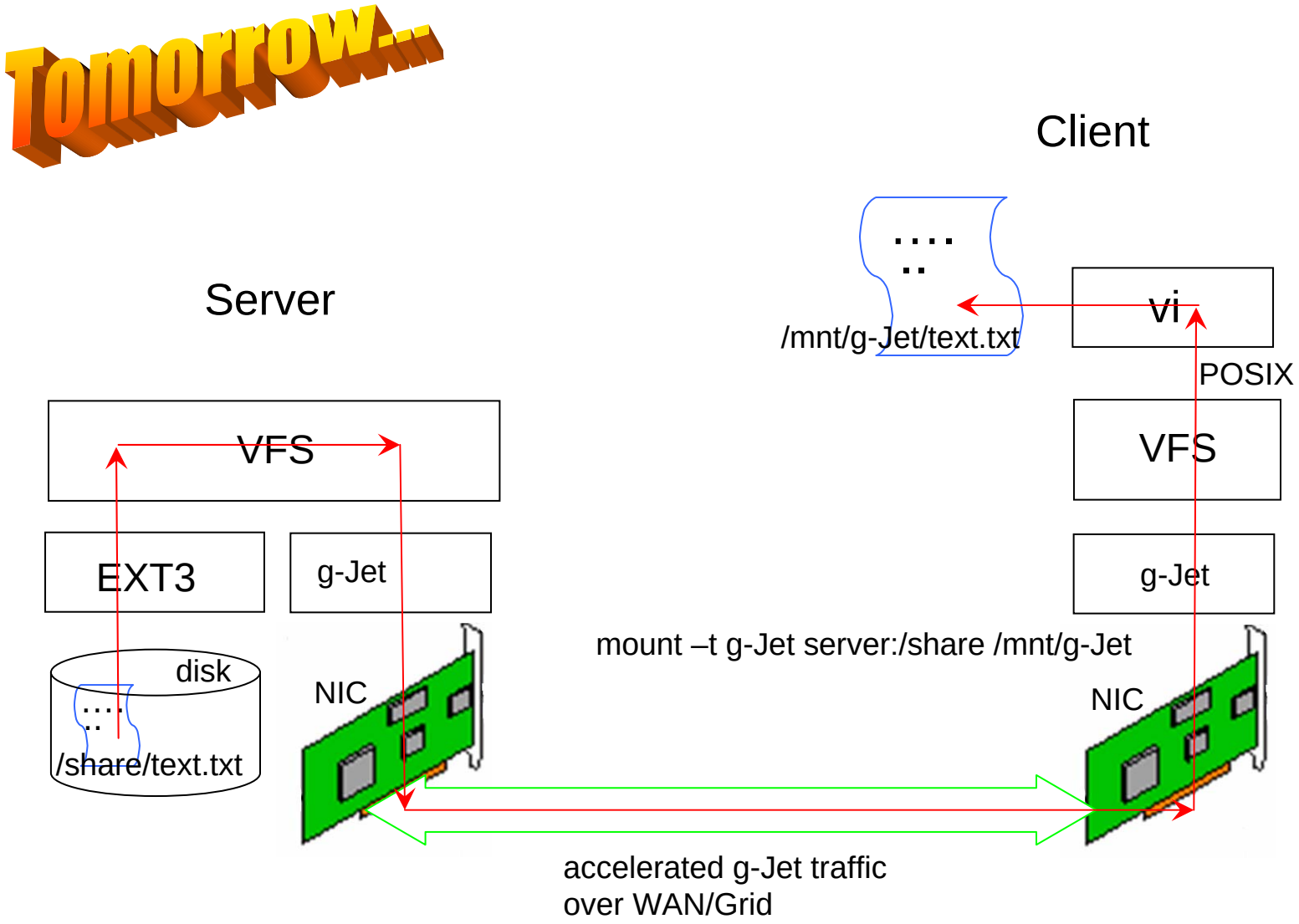
- It is a data communication protocol, a platform, an infrastructure, …

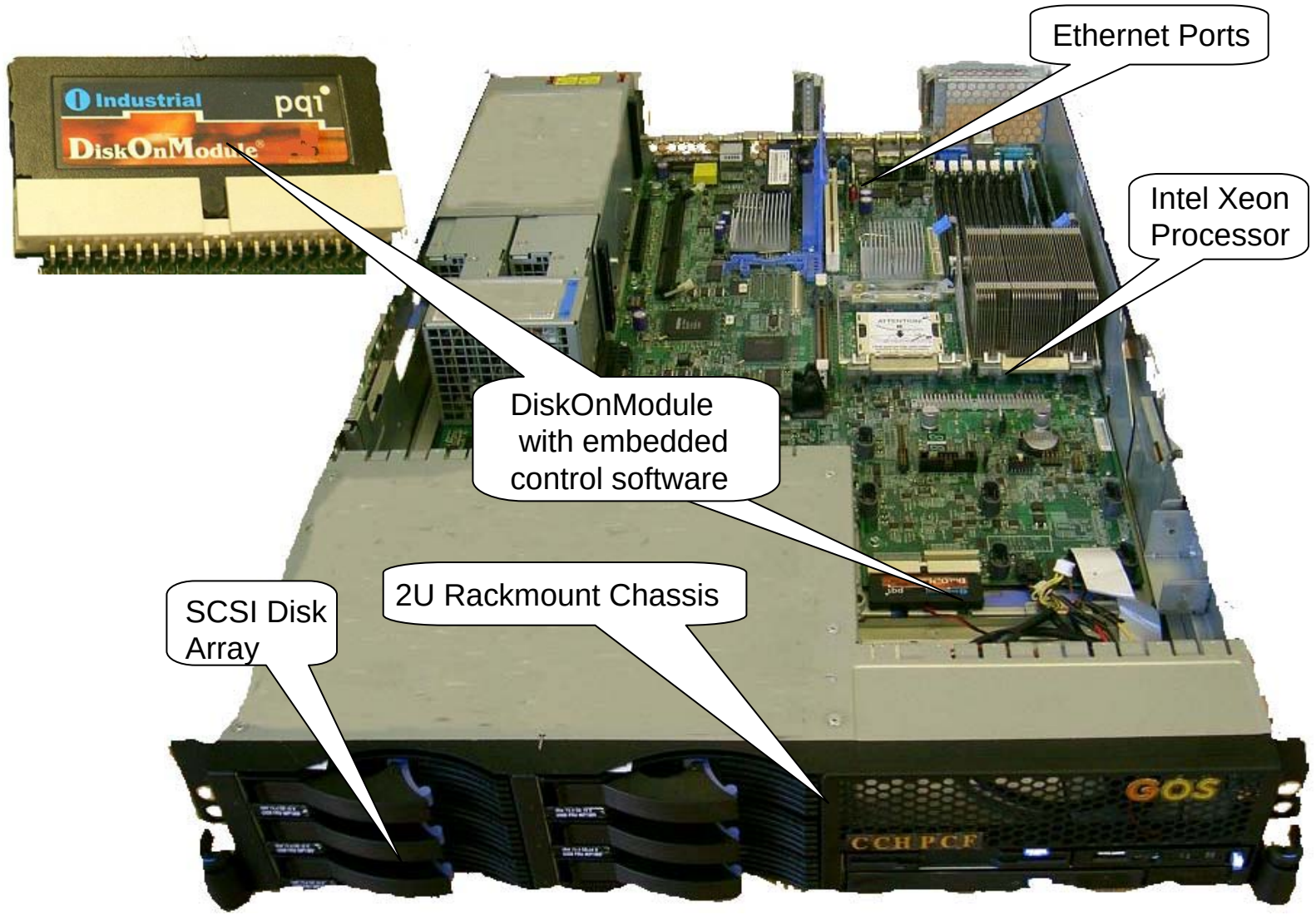
- mount
- ssi


Client

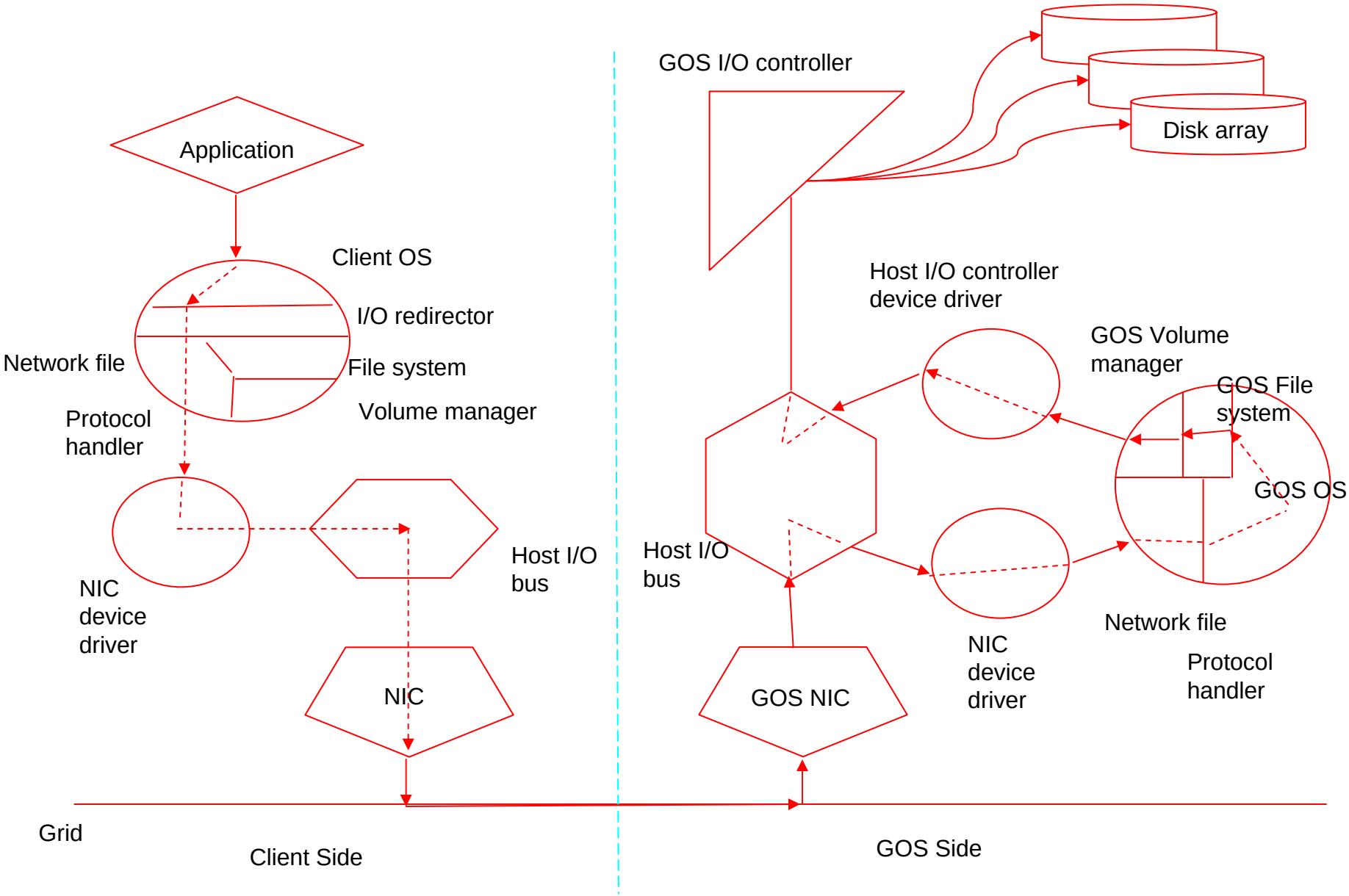
Server



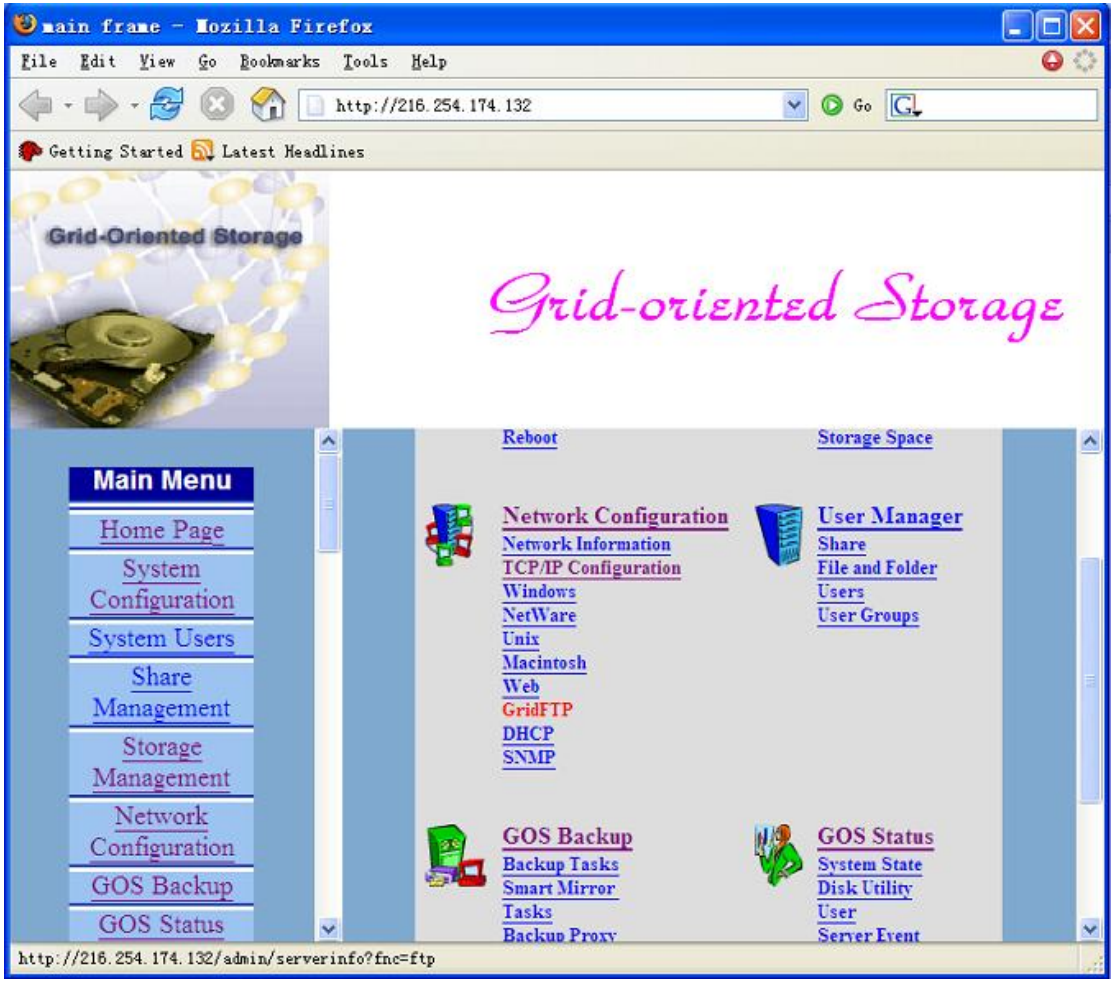






A Grid-oriented Storage (GOS) unit and its DiskOnModule with embedded control software.

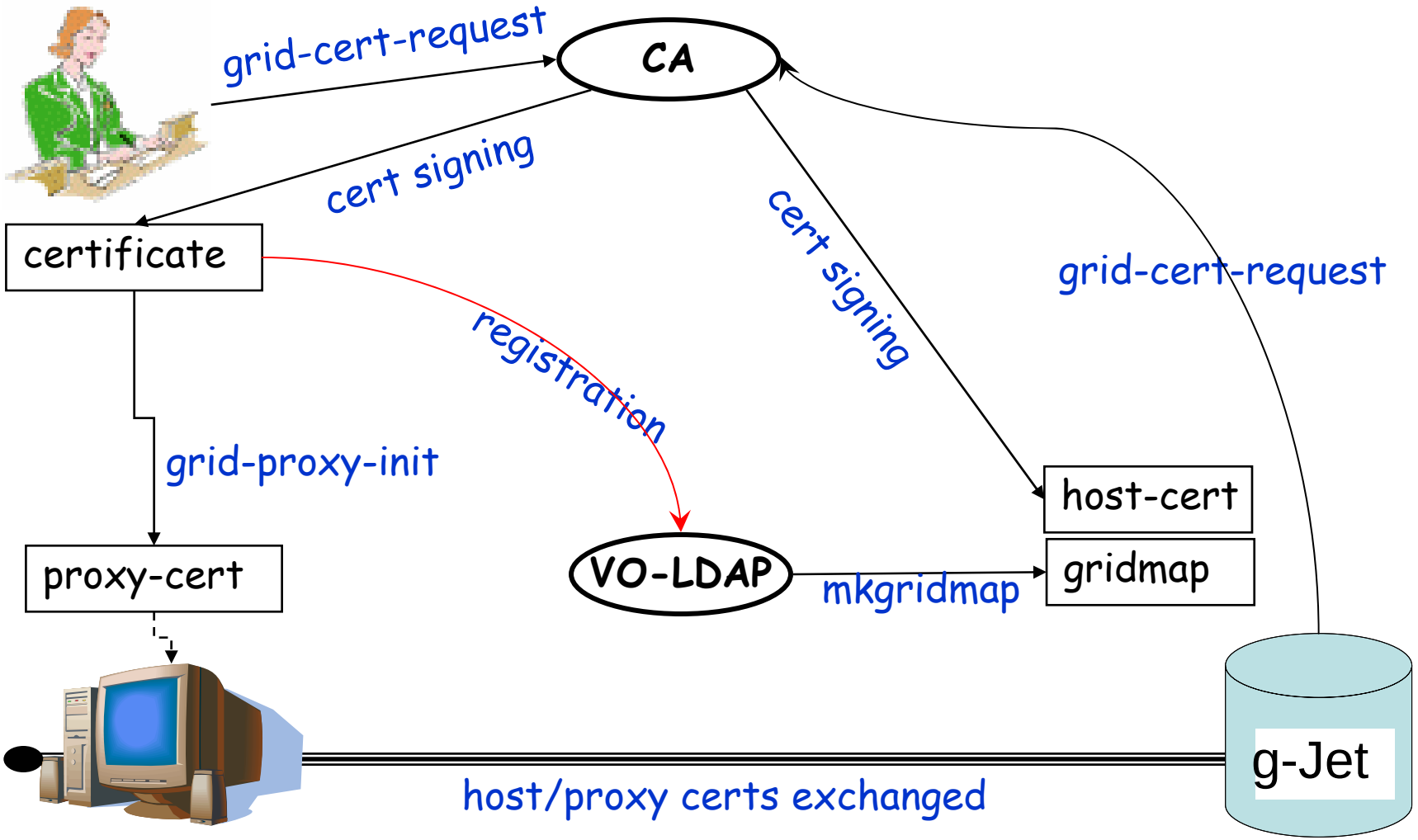



The complete client/GOS path for redirected I/O.



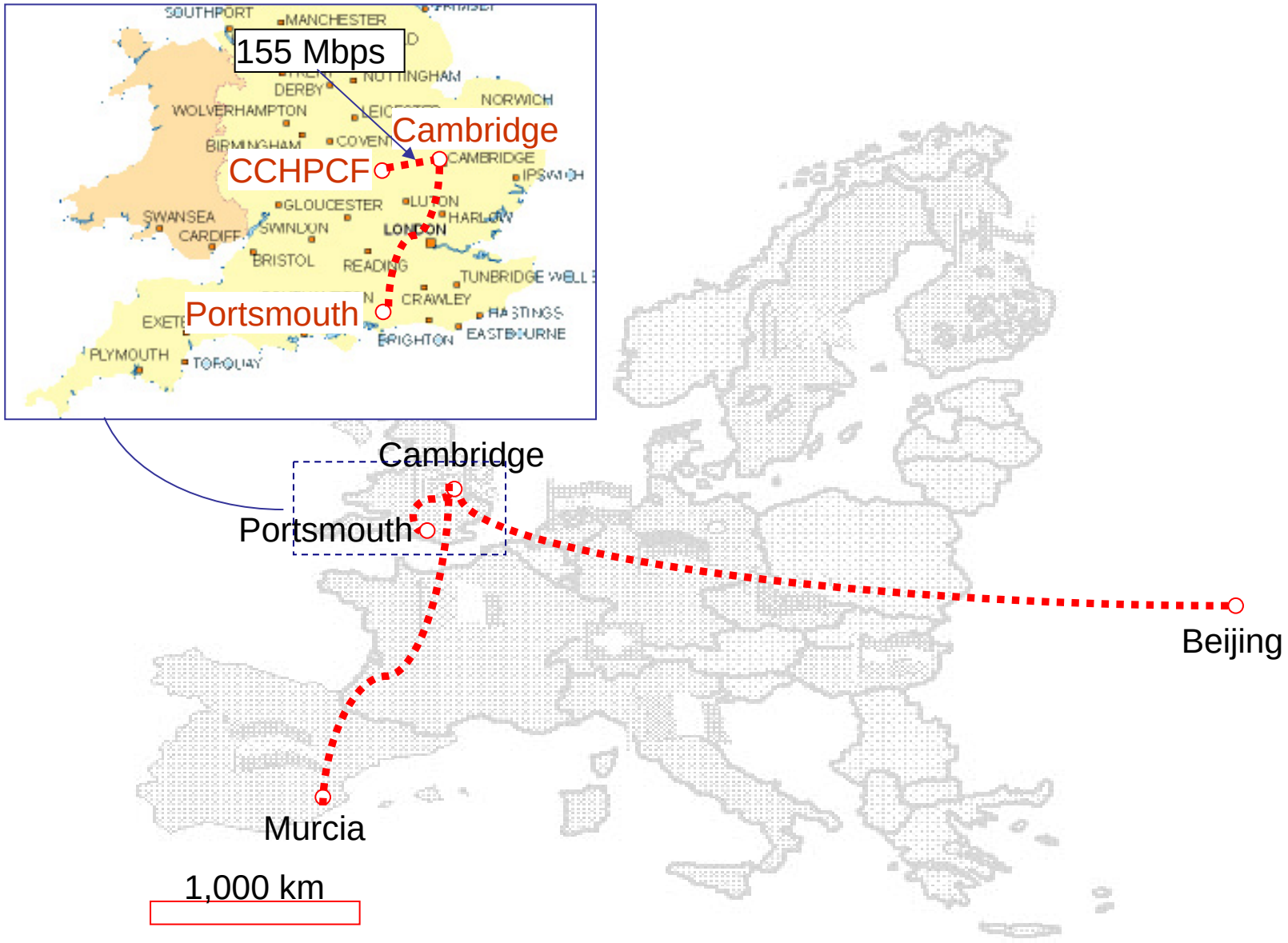

A HTTP-enabled tool has been developed to better view and sort the core performance, service status, and alerts information.

A HTTP-enabled tool has been developed to better view and sort the core performance, service status, and alerts information.



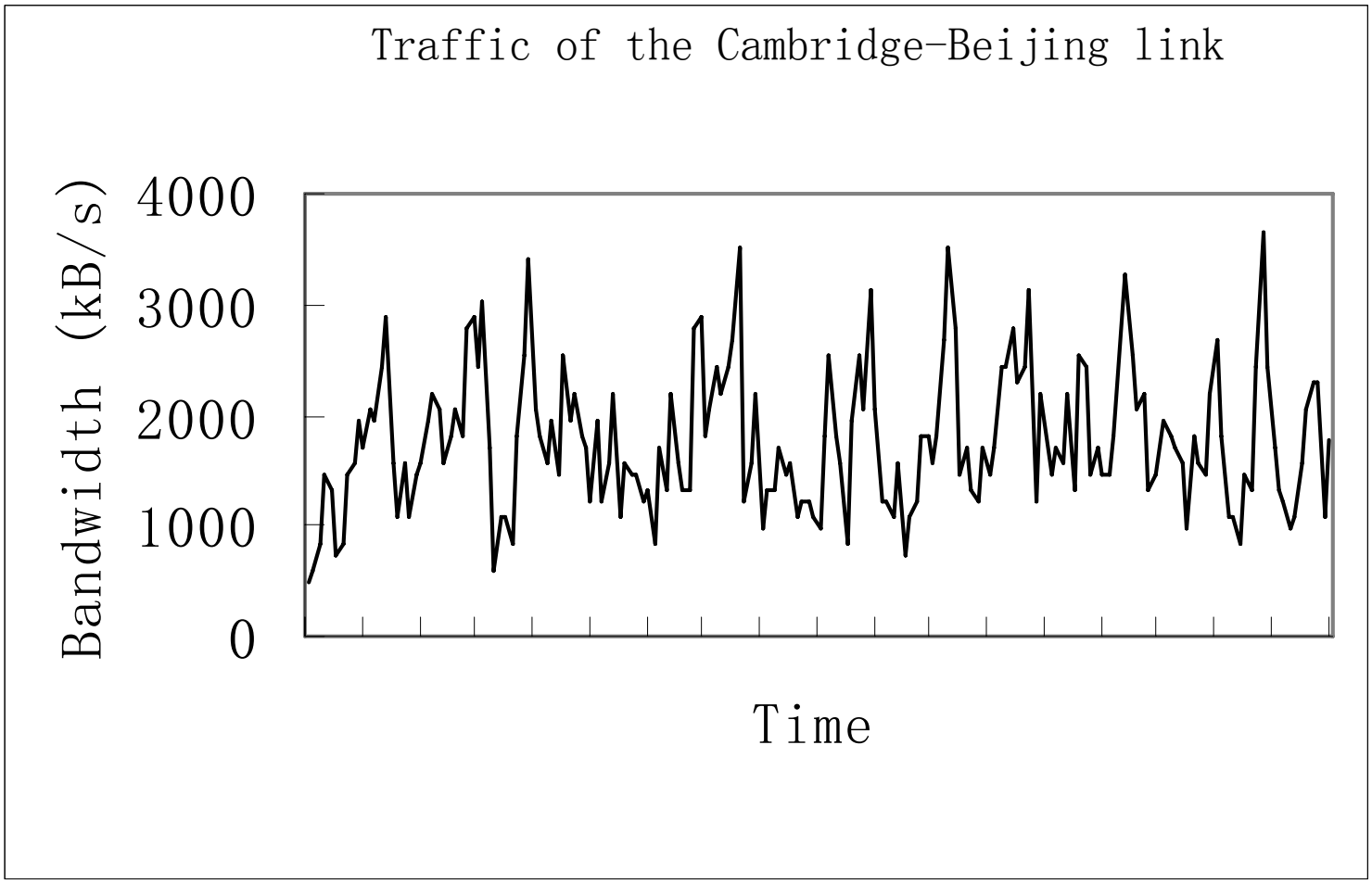
- The source code of the developed gJet is of 40,000 lines in length
- We have spent literally 6 man years in developing and revamping it.

# Real-world Tests



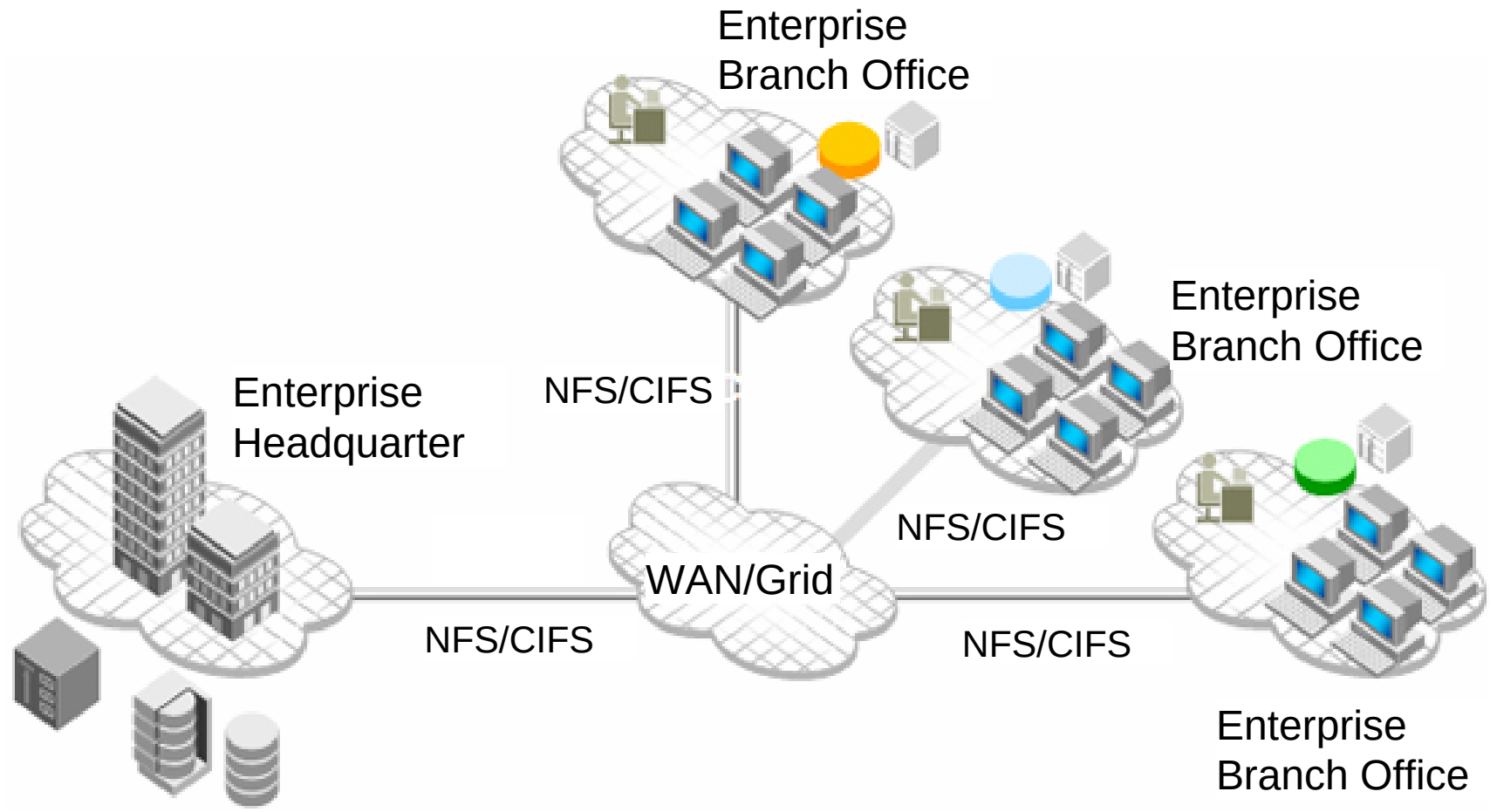

Connectivity map of the EuroAsiaGrid participating sites.

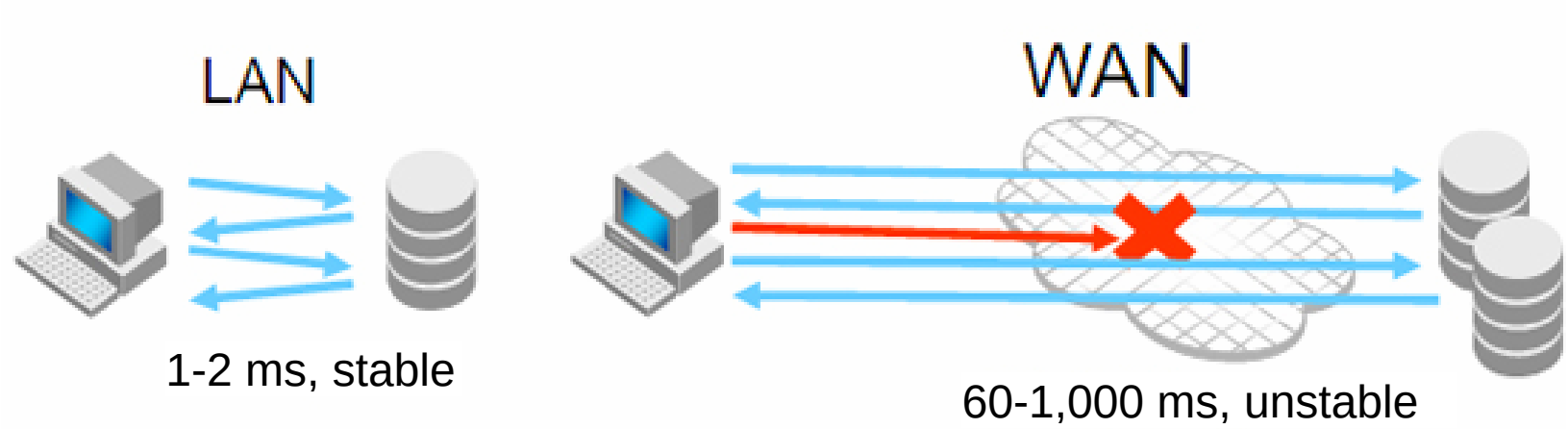
| Link | Router hops | Average RTT (ms) | LAN Connectivity | Client Memory (MB) |
|---|---|---|---|---|
| Cambridge-London | 16 hops | 17 ms | 2 Mbps (London) | 512 MB |
| Cambridge-Portsmouth | 18 hops | 19 ms | 4 Mbps (Portsmouth) | 512 MB |
| Cambridge-Murcia | 23 hops | 62 ms | 10 Mbps (Murcia) | 256MB |
| Cambridge-Beijing | 27 hops | 516 ms | 100 Mbps (Beijing) | 128 MB |



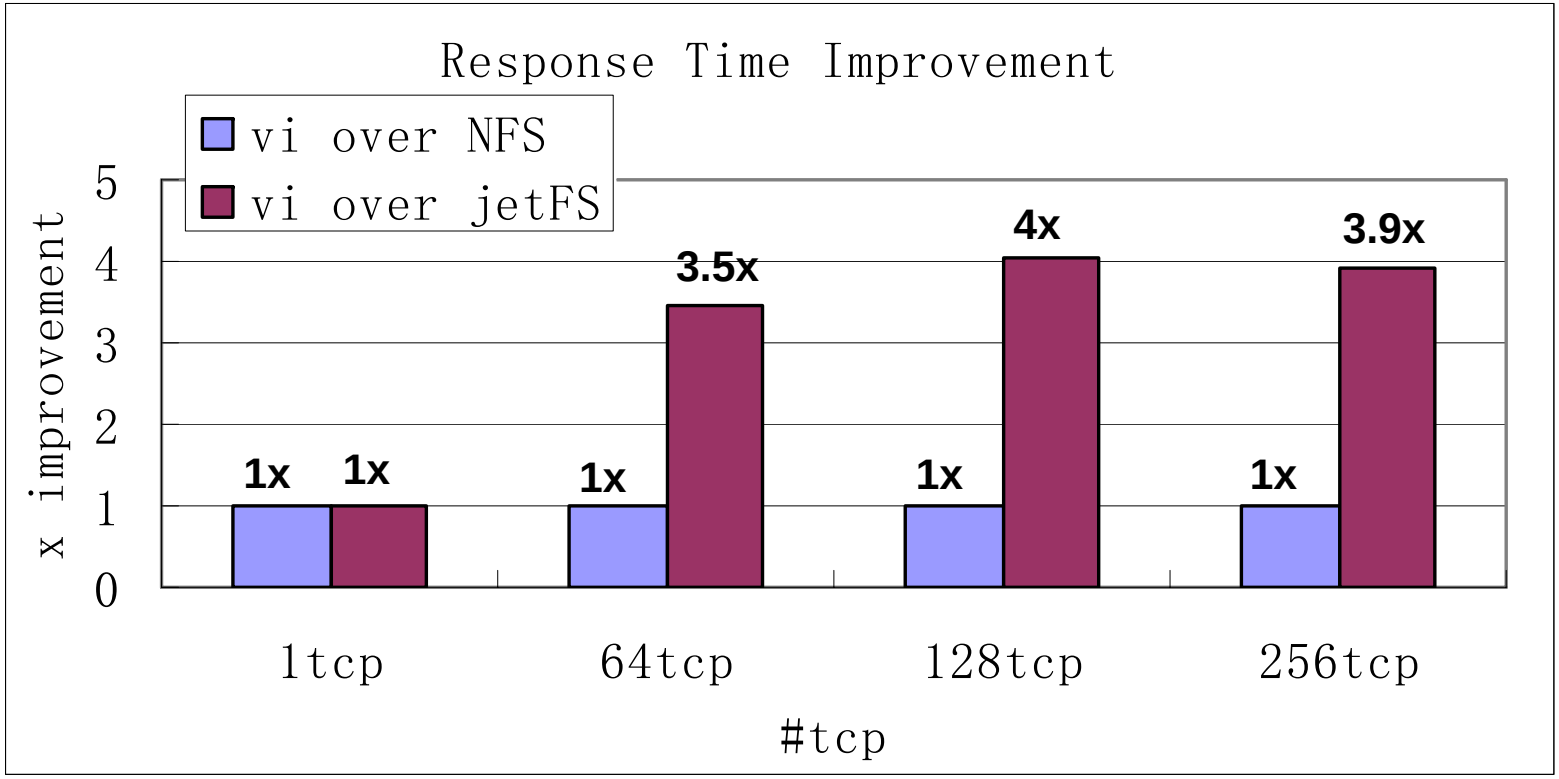

The network fluctuations measured by iperf for 256 simultaneous streams between two workstations, one at Centre for Grid Computing in Cambridge and the other at 3GSports Company in Beijing.

# Accelerating Office applications, vi, …

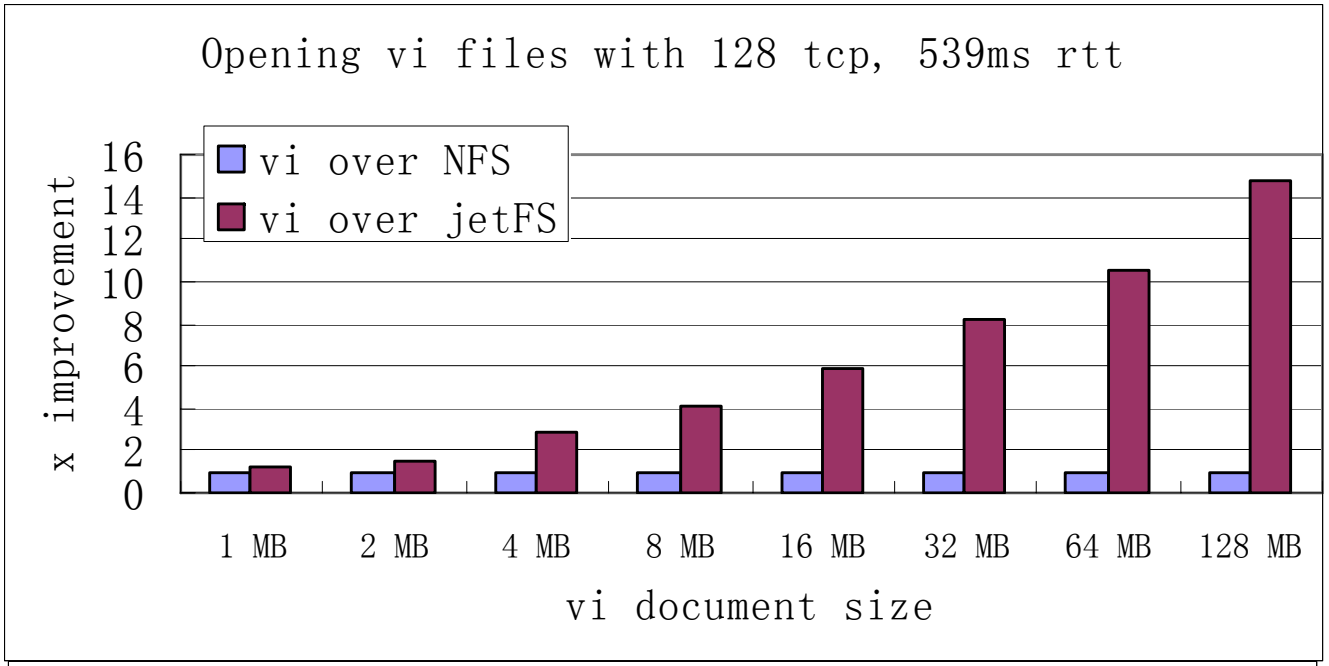


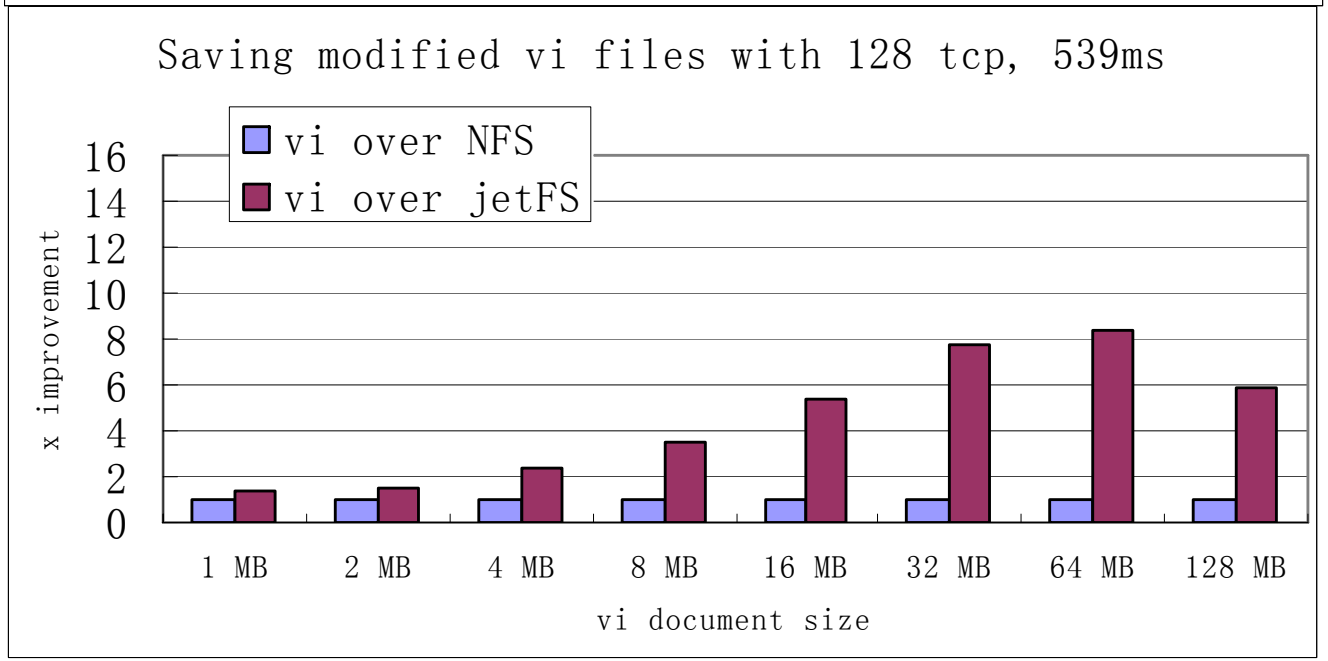


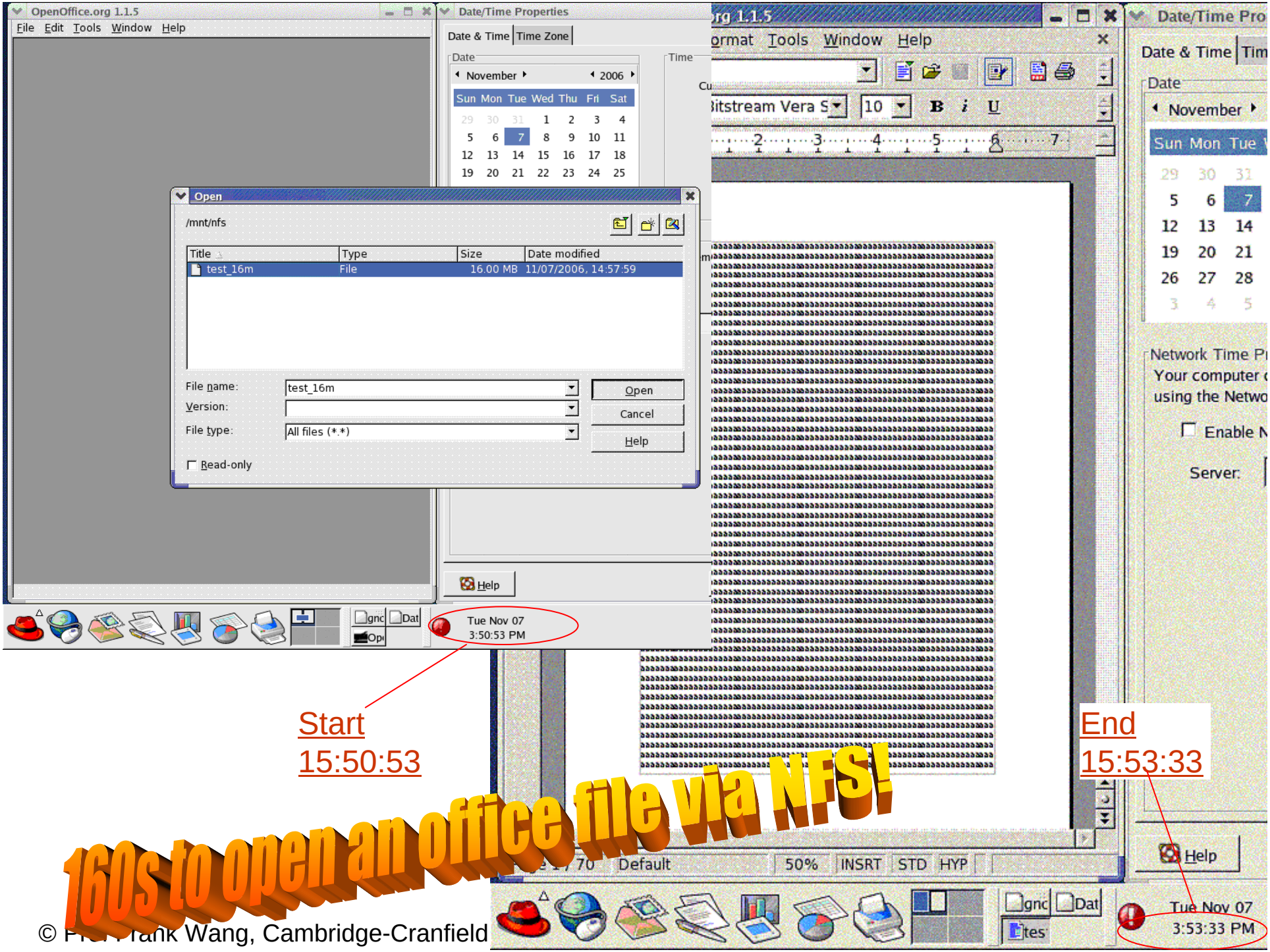
# vi/OpenOffice Writer

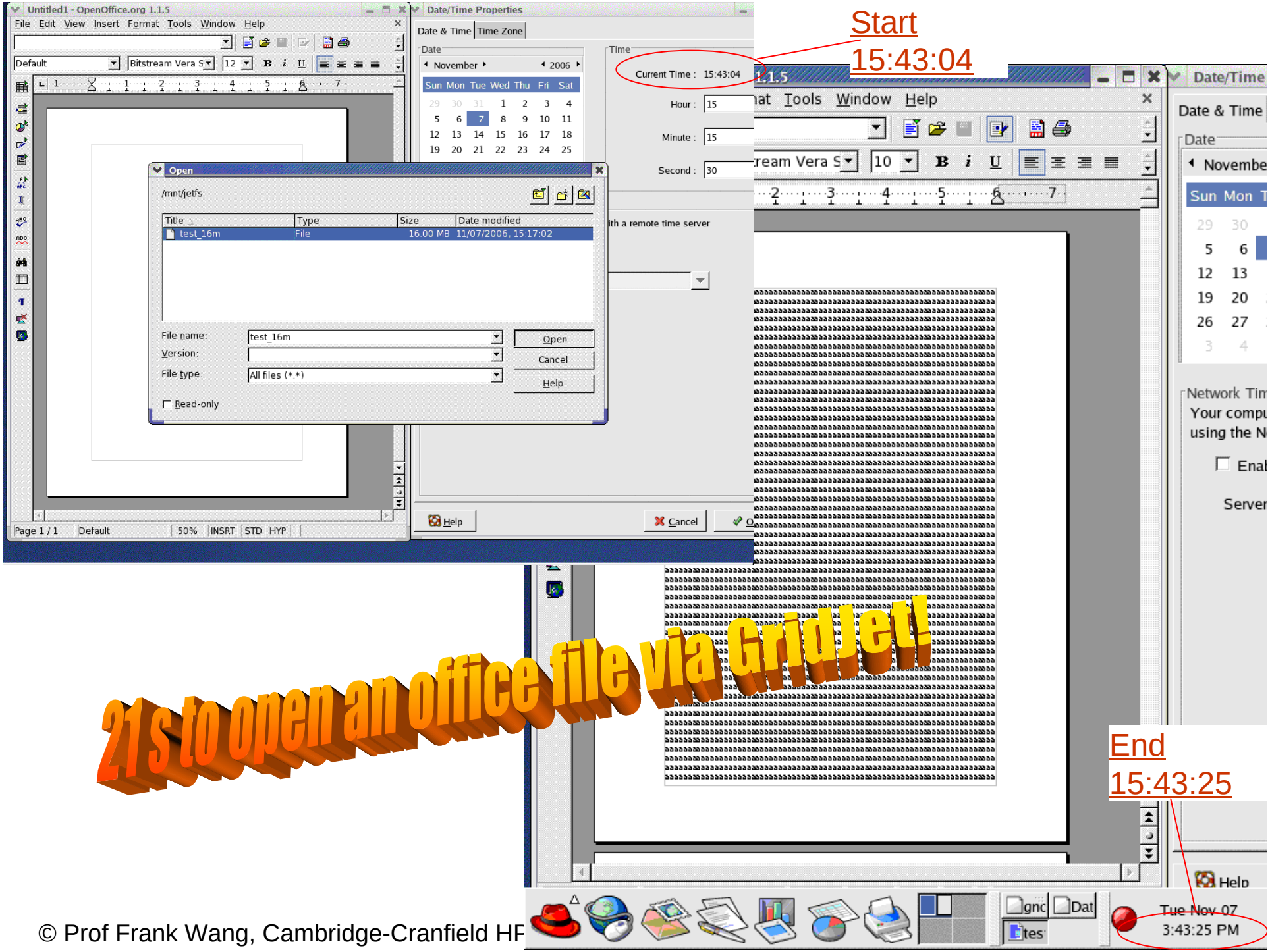



Standard file open of a 8MB vi file over the link with 539ms rtt

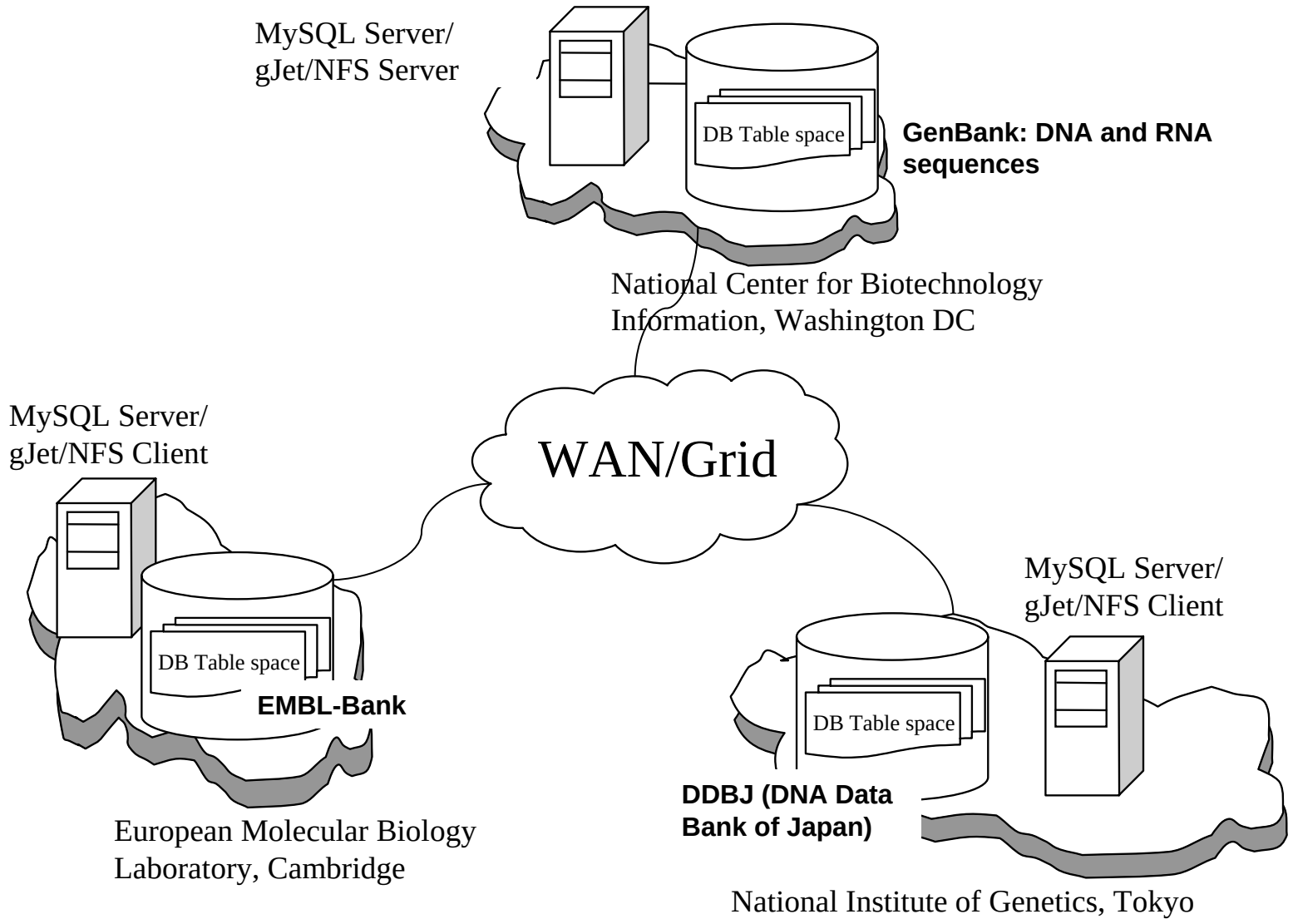






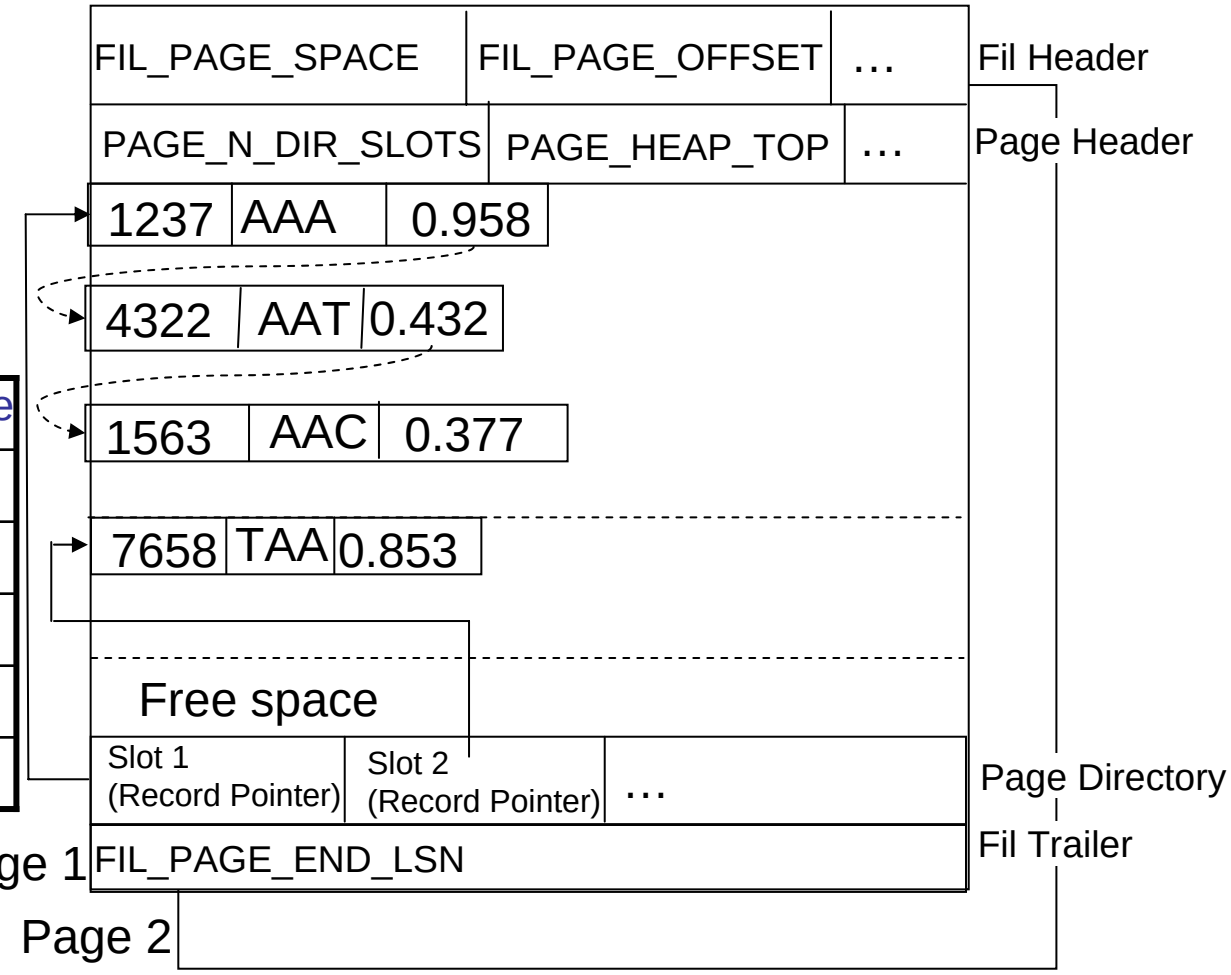
# Accelerating MySQL



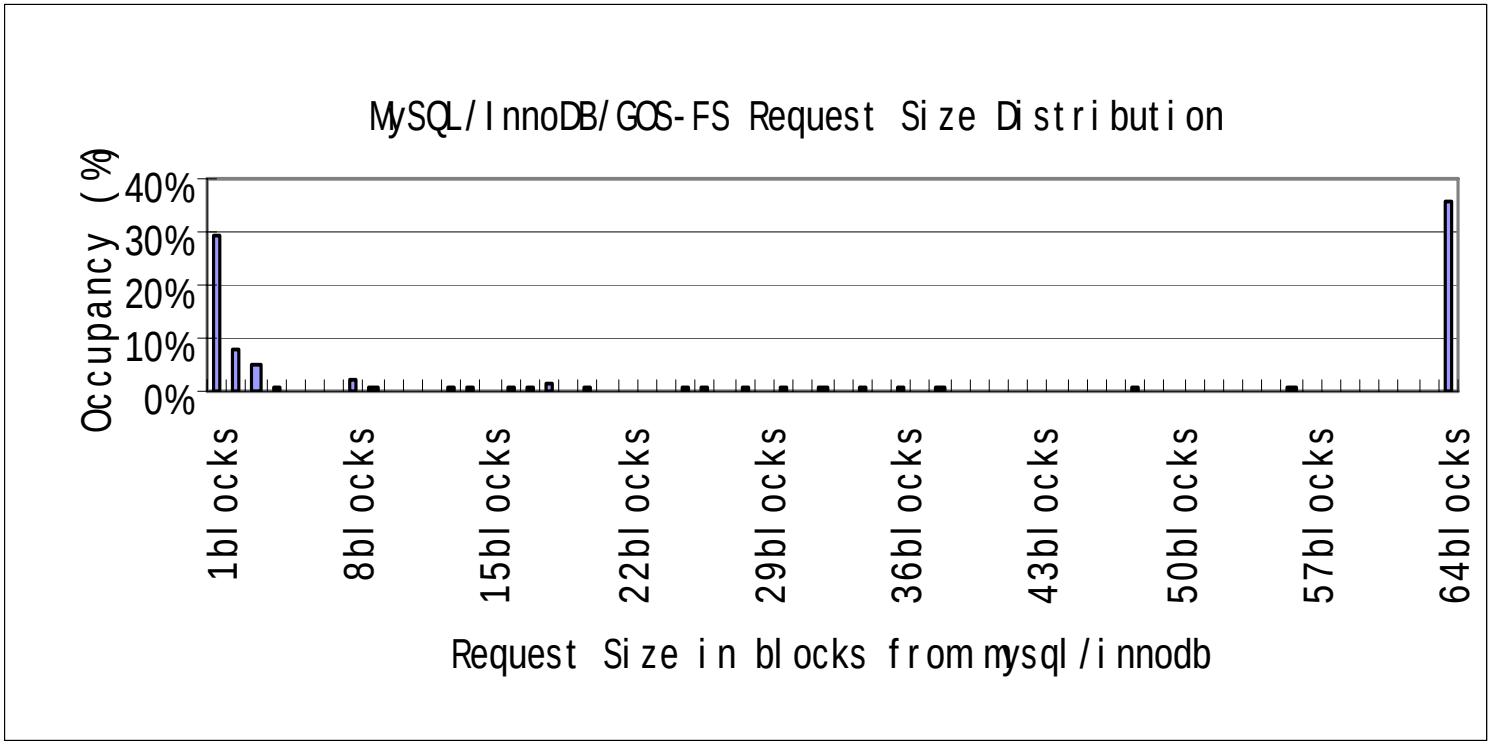
SELECT ID
FROM
Nucleotide_Sequence
where Distance>0.5

Table Nucleotide_Sequence

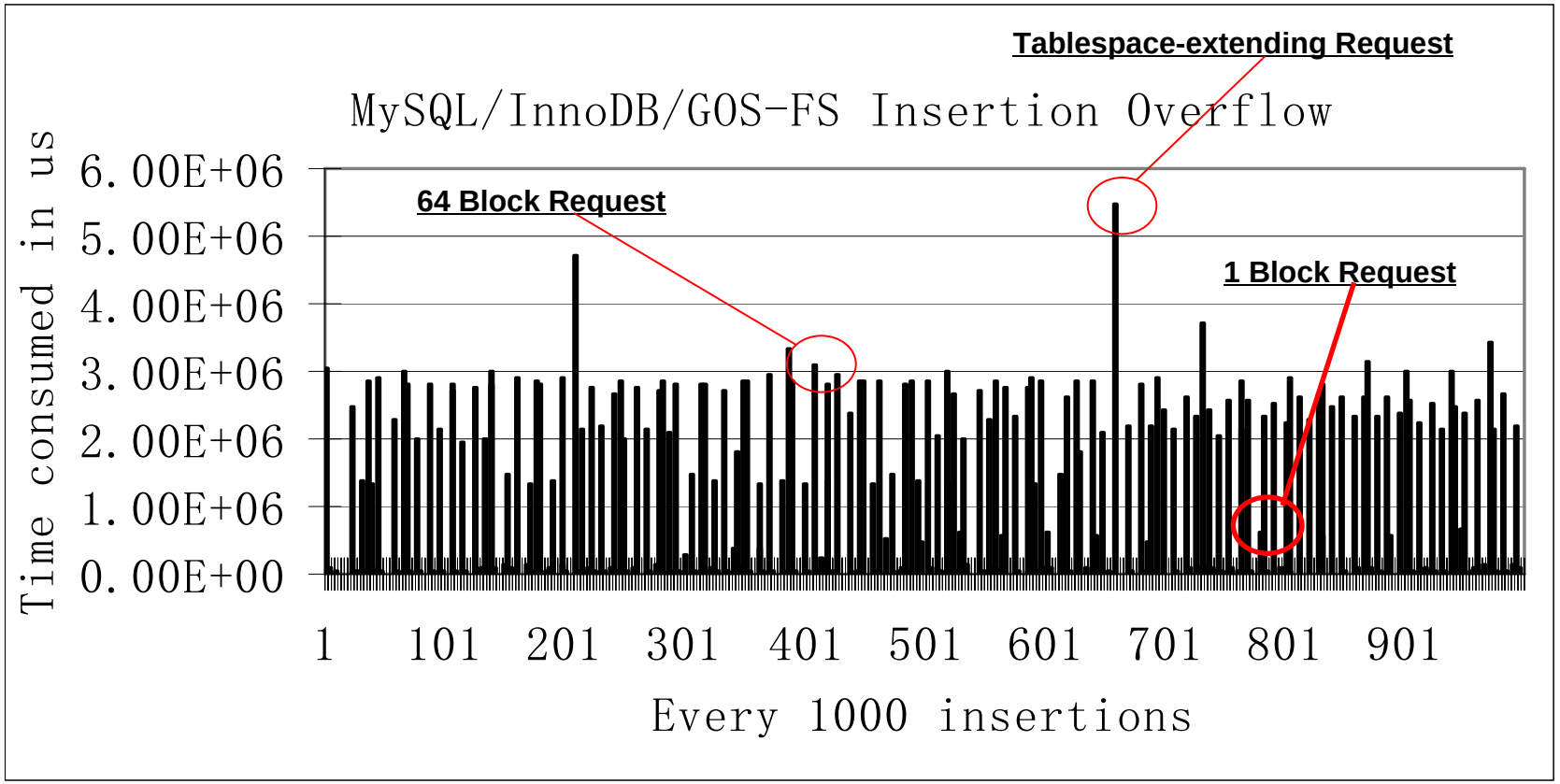
| ID | Sequence | Distance |
|---|---|---|
| 1237 | AAA | 0.958 |
| 4322 | AAT | 0.432 |
| 1563 | AAC | 0.377 |
| 7658 | TAA | 0.853 |
| 2865 | GGG | 0.871 |



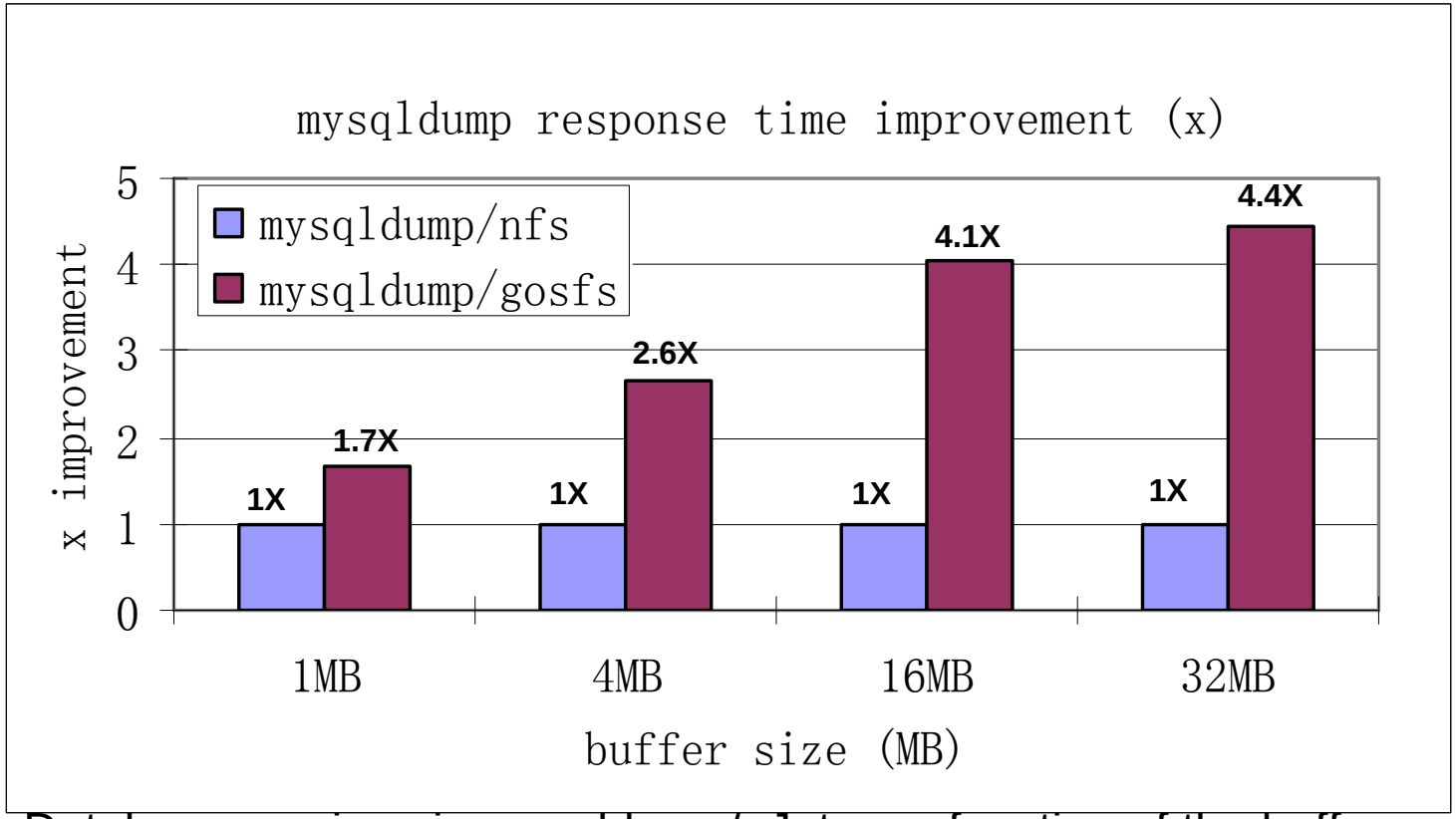

Page structure in the InnoDB.




**MySQL/InnoDB/gJet Request Size Distribution in blocks. 64-block requests occupy nearly 40% of the overall requests whereas 1-block requests occupy around 30%. Insertion requests are issued against a SysBench with 1,000,000 records.**



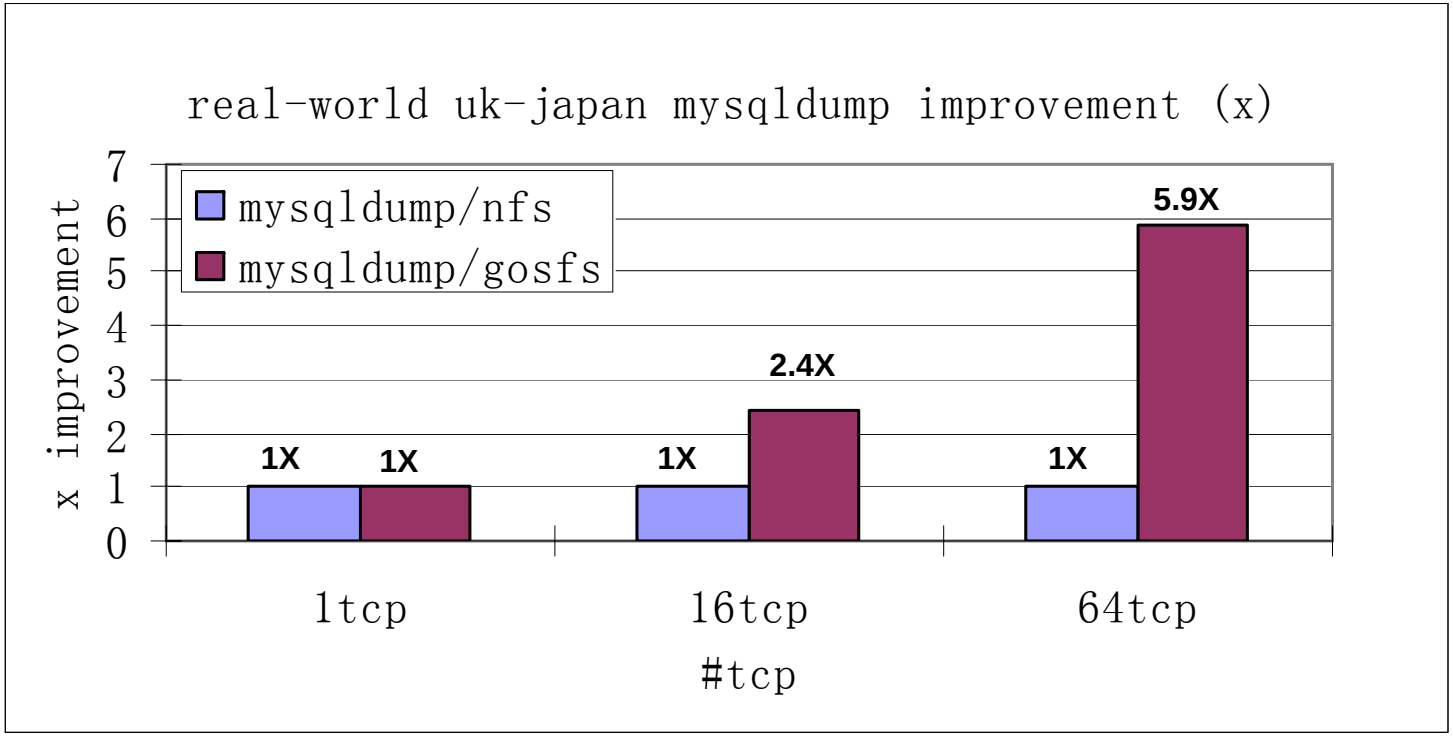

**MySQL/InnoDB/gJet Insertion Flushing. The InnoDB buffer is flushed, represented by a 64-block request if the buffer is filled with contiguous pages , at various intervals depending on memory pressure and I/O activity. Few-block requests are issued from the flushing operations of an InnoDB buffer with dis-contiguous pages. 1-block requests arise from traversing the InnoDB Page B-tree. Most of insertion requests have been satisfied within the InnoDB Buffer, resulting in a very limited amount of response time. Few time-consuming requests can be seen, which extend the table space by 8 MB when the current table space on disk runs out.**



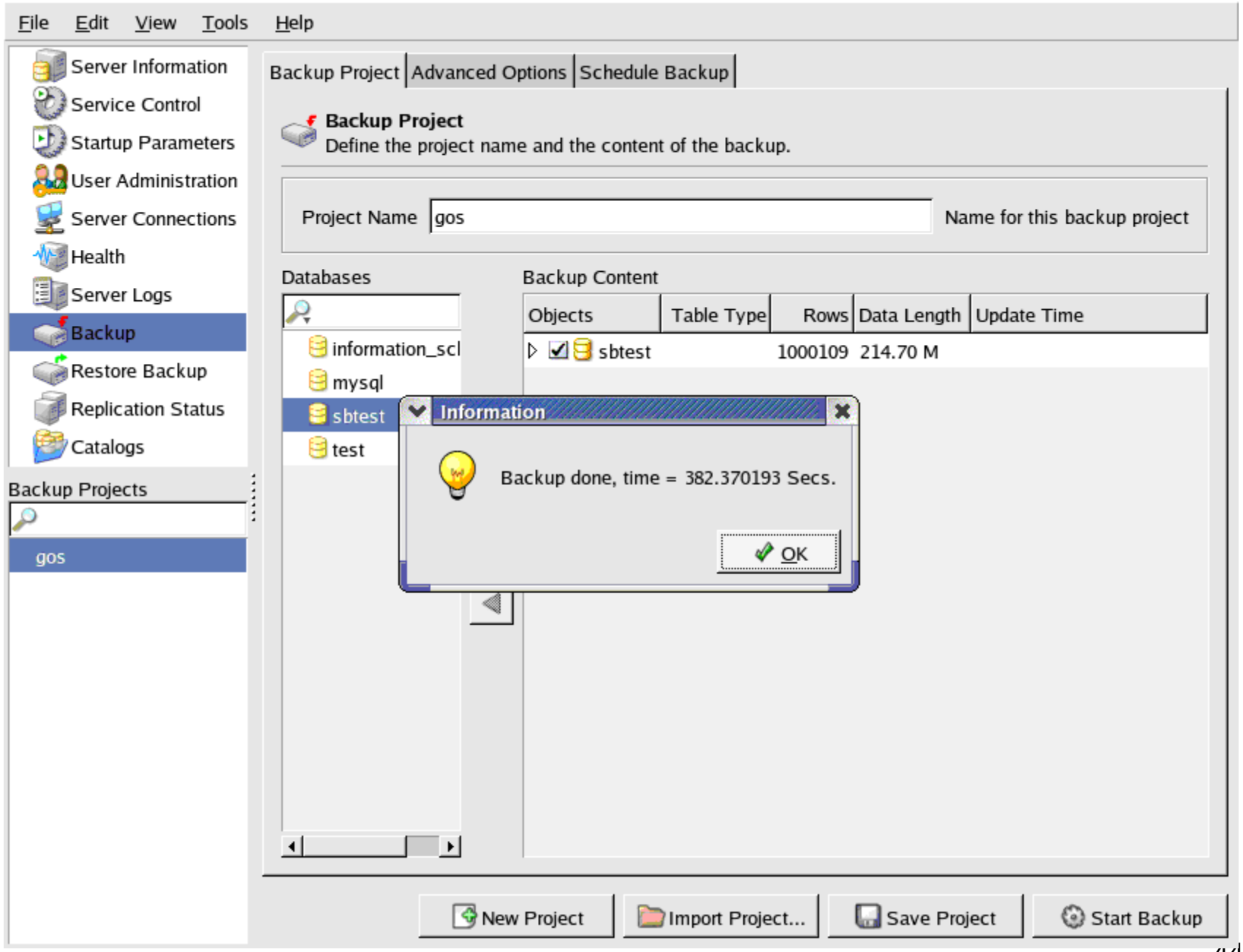

Database moving via mysqldump/gJet as a function of the buffer size over a link with rtt=80ms, #tcp=16. The database contains 1,000,000 records, occupying around 60MB.

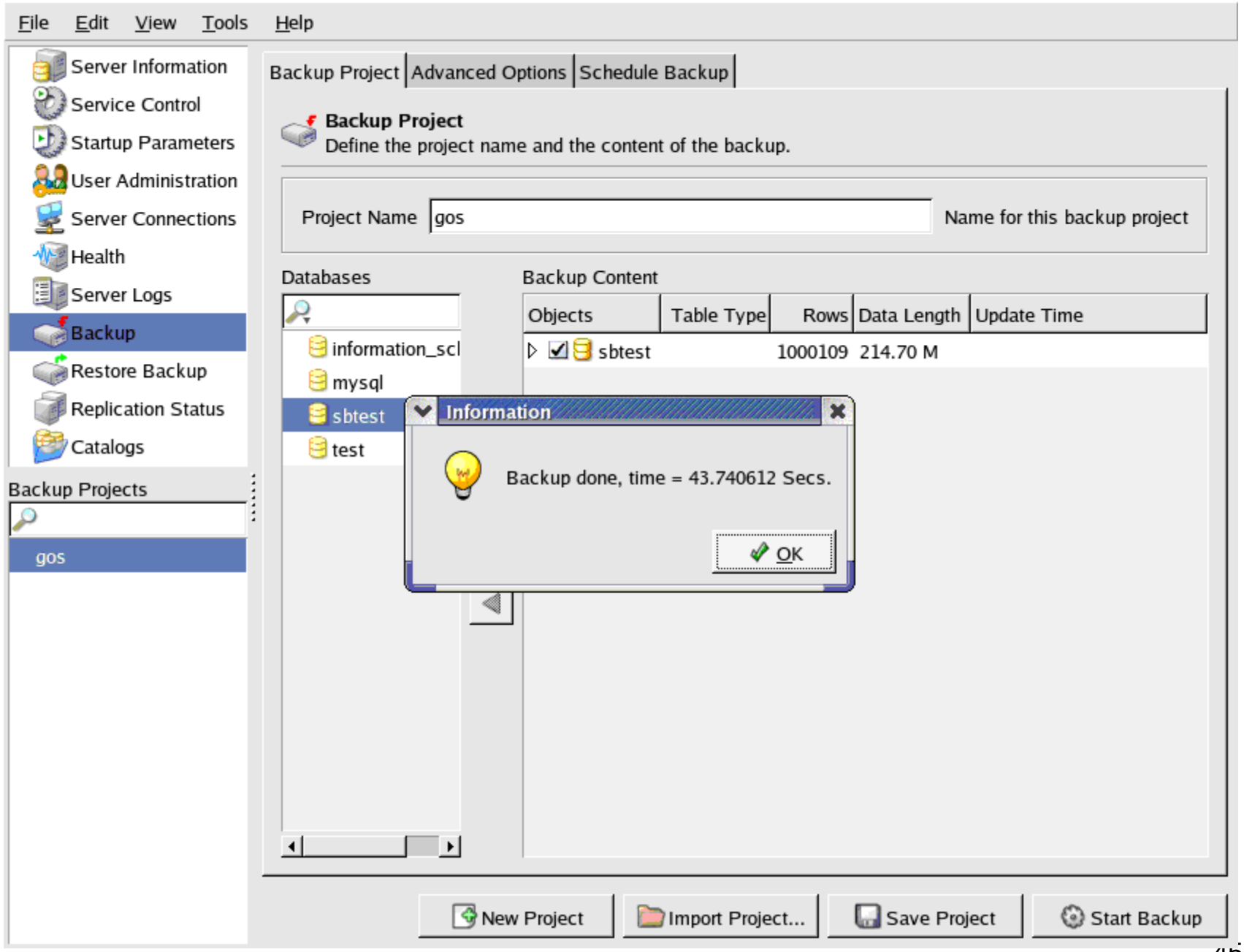
Table 1. Response time (s) of moving a database via mysqldump/gJet. The link is with #tcp=16, buf_size=16MB. The database contains 1,000,000 records, occupying around 60MB.

| rtt(ms) | mysqldump/nfs | mysqldump/gosfs | ftp | gridftp |
|---|---|---|---|---|
| 40ms | 43.8 | 14.4 | 51.8 | 22.9 |
| 80ms | 80 | 19.7 | 92.2 | 32.3 |
| 160ms | 154.2 | 26.4 | 172.2 | 52.9 |
| 320ms | 346.5 | 55.8 | 361.2 | 87.5 |




**Real-world response time improvement as a function of the number of tcp streams (#tcp).**





# Accelerating Firefox

Table 1 Opening time in seconds. RTT=600ms.

| Document size | Firefox/ HTTP | Firefox/ FTP | Firefox/g-Jet (#tcp) | | | |
|---|---|---|---|---|---|---|
| | | | Firefox/g-Jet (1) | Firefox/g-Jet (16) | Firefox/g-Jet (64) | Firefox/g-Jet (128) |
| 256kB | 3 | 3 | 9 | 7 | 4 | 4 |
| 1MB | 11 | 12 | 15 | 13 | 12 | 13 |
| 4MB | 43 | 45 | 62 | 24 | 24 | 24 |
| 16MB | 165 | 175 | 186 | 38 | 31 | 31 |
| 64MB | 324 | 321 | 378 | 66 | 56 | 57 |

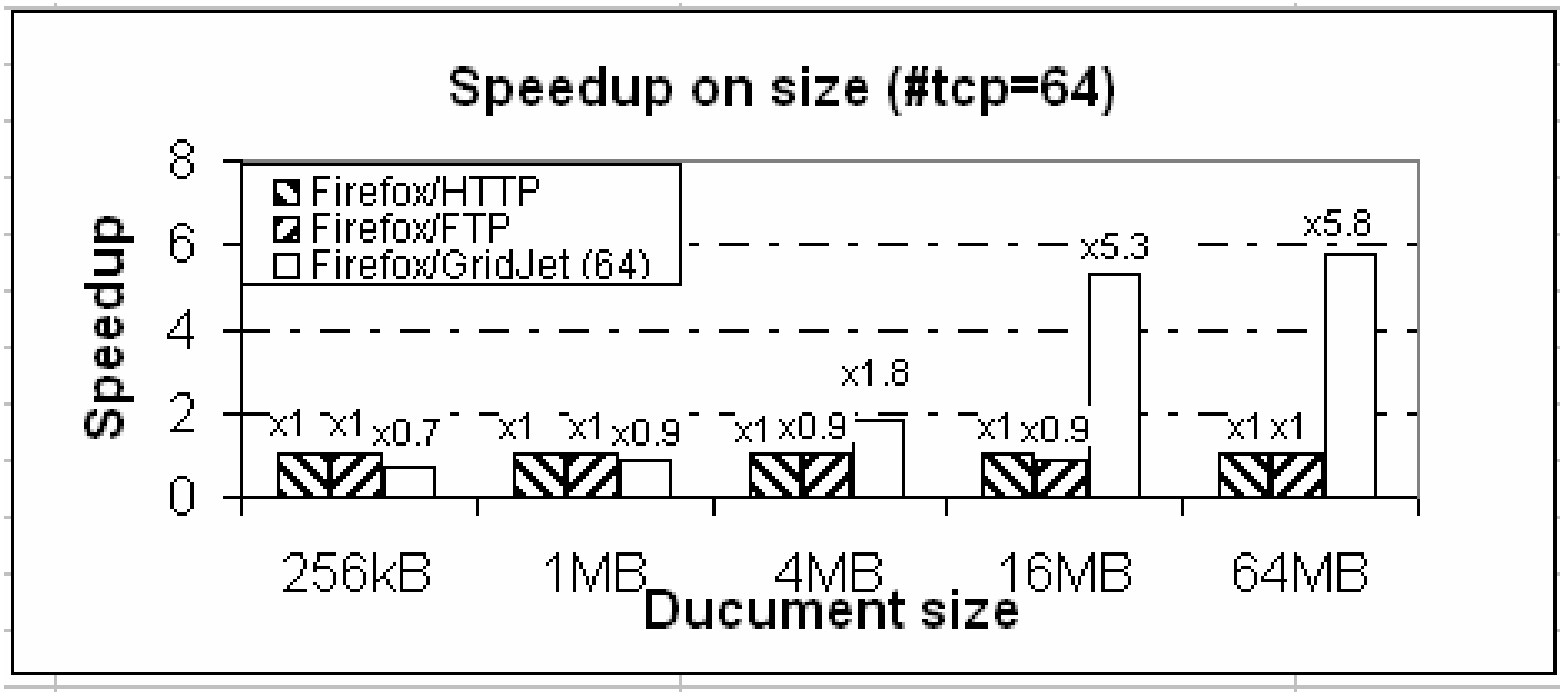



**Fig.6 The test measures the speedup as a function of the document size, namely Firefox over HTTP, FTP and g-Jet (#tcp=64), respectively. Although g-Jet is characterized by an initial transfer time that is larger than the other tools for small documents less than 2MB, this gap tends to decrease as the size of the document increases and for documents larger than 2MB the g-Jet overtakes HTTP and FTP, reaching a speedup of 5.8 when opening a document of 64 MB.**

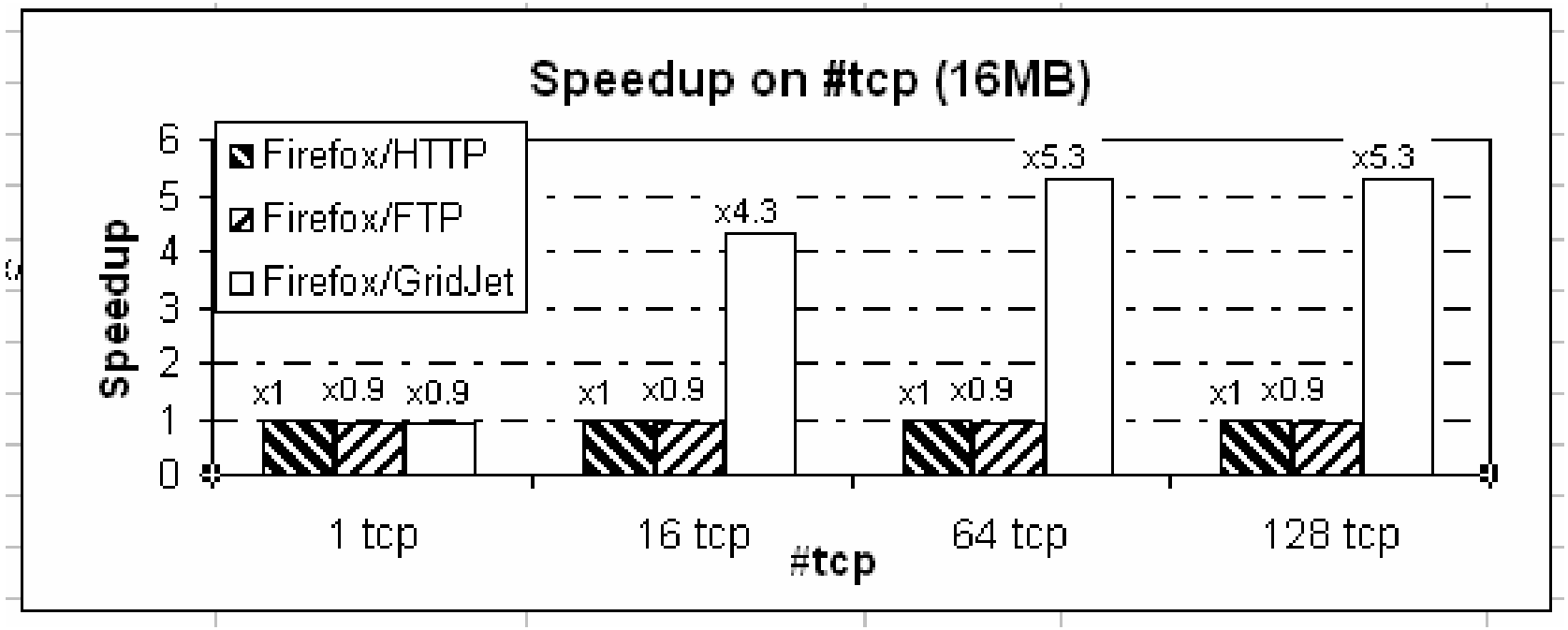



**Fig.7 The test measures the speedup as a function of the number of TCP streams, namely Firefox over HTTP, FTP and g-Jet, respectively. A 16 MB document is used in all the measurements. The maximum improvement goes up to 5.3x with 64 parallel streams. However, g-Jet would eventually consume too much time in managing too many TCP streams, slowing down such an improvement with further increased #tcp.**

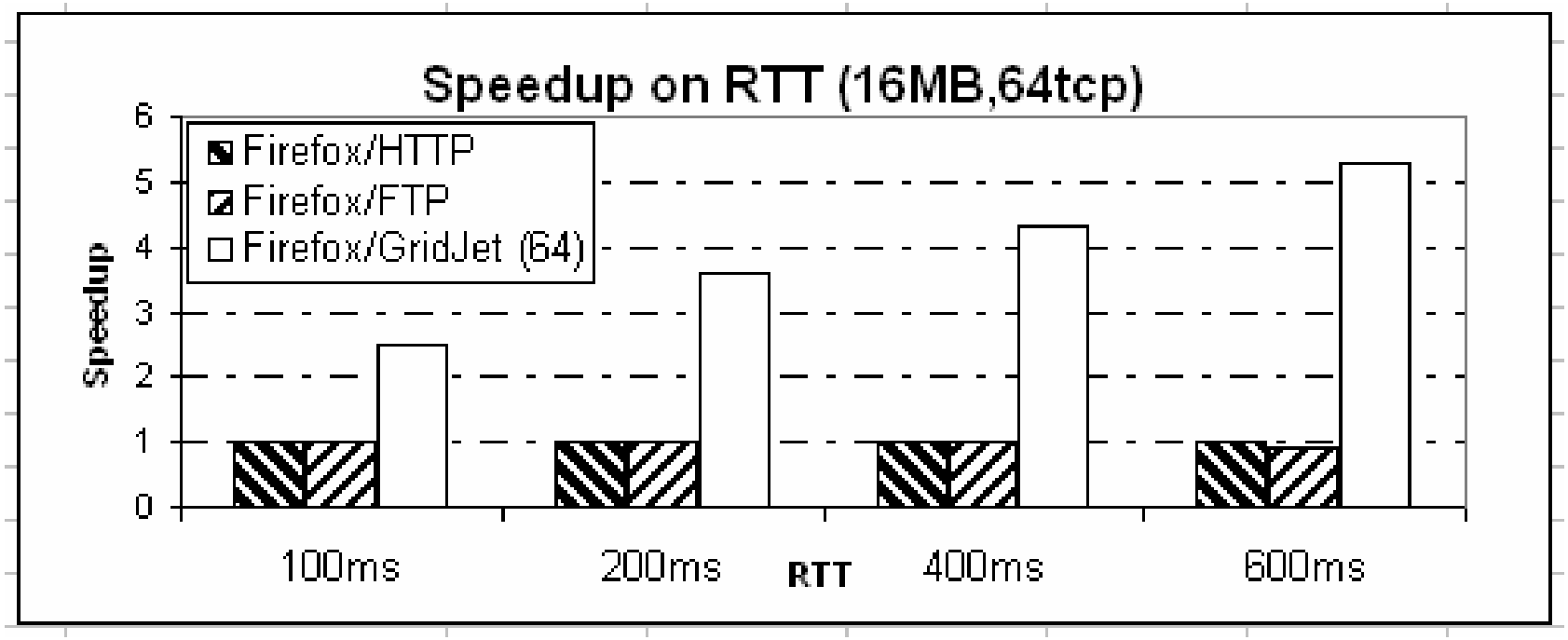



**Fig.8 The test measures the speedup as a function of the Round Trip Time (RTT), namely Firefox over HTTP, FTP and g-Jet (#tcp=64), respectively. A 16 MB document is used in all the measurements. In the figure, g-Jet outperforms the classic HTTP by factors ranging from 2 to 6, and this gap widens when RTT is increased further (RTTs of 60-1000 ms typify WWW environments).**

# Accelerating Media Players

- Media players are well-used to access and share those documents in audio and video formats either on wired Internets or on wireless Local Area Networks.
- A media player is normally featured with multimedia support, skins/themes, plugins/extensions, portable device compatibility including cellular phones, PDA and ipod, etc.
- However the cross-platform media access via conventional HTTP often results in long waits.

# GridJet provides a LAN-like performance!

| file_size = 36,411,356 Bytes | | number of frames = 6150 | | |
|---|---|---|---|---|
| network: 100Mbps | start_time (s) | play_time (s) | drop_frame | frame rate (fps) |
| ext2(local) | 1 | 209 | 0 | 29.4 |
| gjet (rtt=600ms) | 65 | 229 | 12 | 26.9 |
| nfs (rtt=600ms) | 25 | 392 | 173 | 15.7 |
| ftp (rtt=600ms) | 73 | 404 | 175 | 15.2 |
| http (rtt=600ms) | 8 | 402 | 179 | 15.3 |

| file_size = 36,411,356 Bytes | | number of frames = 6150 | | |
|---|---|---|---|---|
| network: 10Mbps | start_time (s) | play_time (s) | drop_frame | frame rate (fps) |
| ext2(local) | 1 | 209 | 0 | 29.4 |
| gjet (rtt=600ms) | 56 | 245 | 24 | 25.1 |
| nfs (rtt=600ms) | 26 | 397 | 181 | 15.5 |
| ftp (rtt=600ms) | 66 | 409 | 177 | 15.0 |
| http (rtt=600ms) | 7 | 405 | 179 | 15.2 |

# Google Earth…

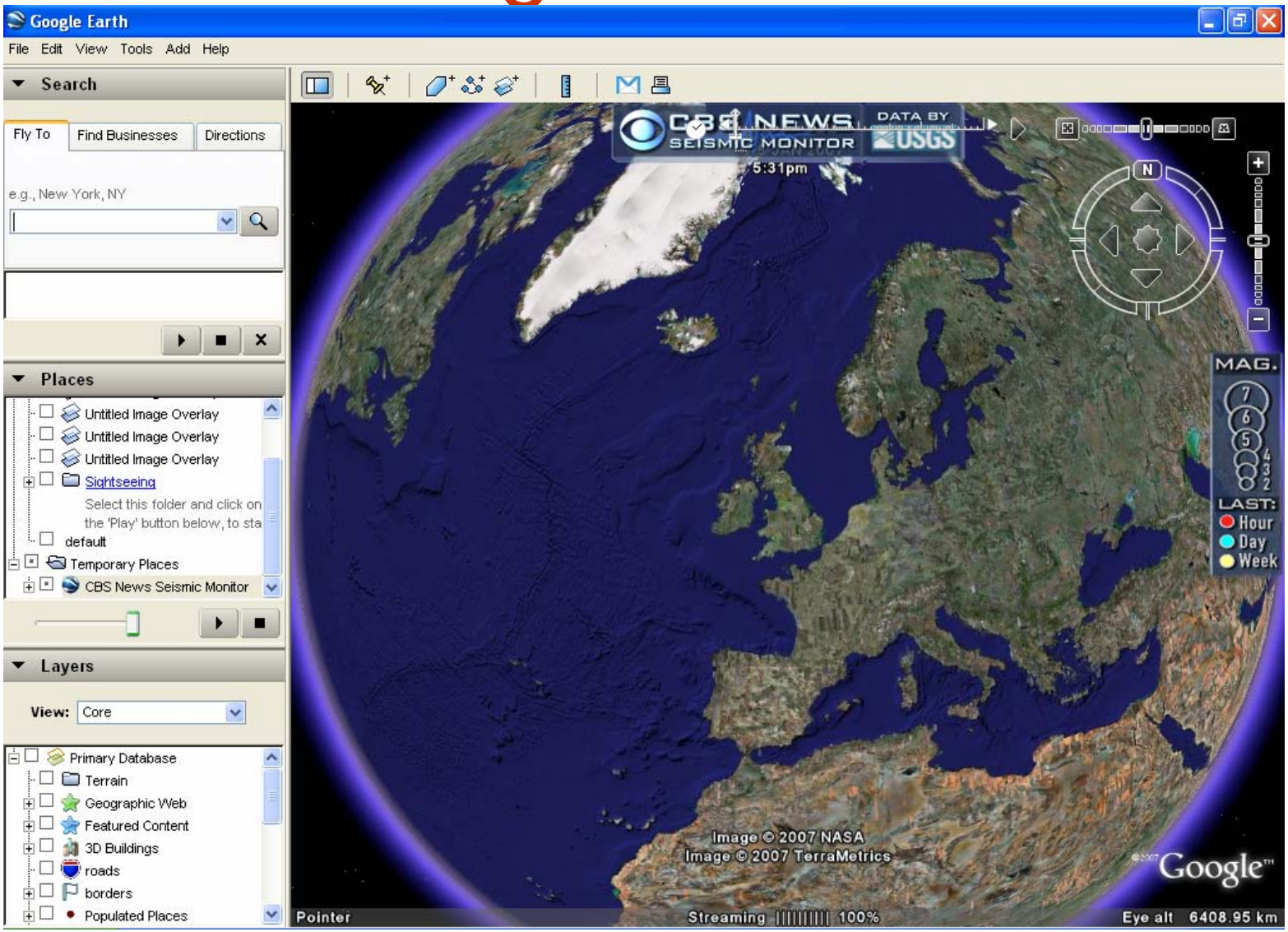


Table Google Earth performance in a real-world test between Cambridge and Beijing.

| | access time in seconds, rtt=539ms | | Speedup |
|---|---|---|---|
| #tcp | GoogleEarth/g-Jet | GoogleEarth/HTTP | |
| 1tcp | 448 | 400 | 0.9 |
| 16tcp | 73 | 400 | 5.5 |
| 64tcp | 52 | 400 | 7.7 |
| 128tcp | 51 | 400 | 7.8 |

# Checklist

- We were inspired by GridFTP that uses multi-streams and GSI
- We have been attempting to implement a network/grid filesystem in a similar way
- Network file system is not a (disk) filesystem
- It is a data communication protocol, a platform, an infrastructure
- Other (distributed) applications can be deployed on it, like db, vi, firefox, mplayer, Google Earth…
- Mount, ssi

# Conclusion

- It is a simple technique
- It worked!

# GOS dissemination activities

Since January 2007, the GOS invention led to invitations to present talks at

- Princeton University (15 January 2007, hosted by Prof Kai Li)
- Cambridge University (Computer Laboratory, 25 January 2007, hosted by Prof Jon Crowcroft)
- Rolls Royce (23 March 2007, hosted by Prof Ric Parker, Director of RR Research & Technology)
- BBC (29mar07, DSNetUK, hosted by Prof Wright, BBC Future Media & Technology, Surrey)
- Xerox (15apr07, Webster, NY, USA, hosted by Dr. Sophie Vandebroek, President of Xerox Technology Innovation Centre)
- Carnegie Mellon University (17Apr07, Pittsburgh, USA, hosted by Prof Gibson)
- CERN (to be presented in 2007, hosted by Dr. Flavia Donno, CERN/IT-GD)
- Manchester University (to be presented in 2007, hosted by Dr. Jim Miles)
- Oxford University (Computer Laboratory, to be presented in 2007, hosted by Dr. Andrew Simpson)
- etc.

# Comments from independent reviewers:

- This proposed work has all the potentials of keeping the UK research and development – world class.
- It is certainly first of its kind
- The academic partners are well-experienced in the field and are known to be serious researcher
- this has the potential to change how networks and applications associated with networks are utilised.

# Comments from independent reviewers (cont.):

- The idea of a storage system that is optimized for grid protocols is a potential winner.
- The system can provide high sustainable throughput for long distant data transfer, specifically optimized for Grid computing environment.

# About GOS …

"This is a rather famous follow-on project :-)"

--- Prof. Garth Gibson, Father of NAS and NASD, Carnegie Mellon University

## GARTH GIBSON


**Associate Professor**
Computer Science Dept and
Dept of Electrical and Computer Engineering
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213-3891
E-mail: garth.gibson@cs.cmu.edu

and

**Co-Founder and Chief Technology Officer**
Panasas, Inc. www.panasas.com
1501 Reedsdale Street, Pittsburgh PA 15233
Phone: 412-323-3500, FAX: 412-323-3511
E-mail: garth.gibson@panasas.com



G. Gibson, D. Nagle, K. Amiri, J. Butler, F. Chang, H. Gobioff, C. Hardin, E. Riedel, D. Rochberg, J. Zelenka, A Cost-Effective, High-Bandwidth Storage Architecture, Proceedings of the 8th Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 1998.

G. Gibson and R. Van Meter, Network Attached Storage Architecture, Communications of the ACM, Vol. 43, No. 11, 2000

SDI/LCS

**Seminar Series**

PARALLEL DATA LAB

**DATE:** Thursday, May 17 , 2007
**TIME:** 12:00 pm - 1:00 pm
**PLACE:** Wean Hall 8220

**SPEAKER:**

**Frank Wang**
Professor, Chair in E-Science and Grid Computing, Director of Centre For Grid Computing
Cambridge-Cranfield High Performance Computing Facility (CCHPCF)

**TITLE:**

## Grid-oriented Storage: a successor to network-attached storage?

**ABSTRACT:**

Prof Wang and his group have developed a prototype Grid-oriented Storage (GOS) appliance, which is expected to be a successor to network-attached storage (NAS) in the Grid era. A GOS-specific File System (GOS-FS), the single-purpose intent of a GOS OS, and secure interfaces via Grid Security Infrastructure (GSI) motivate and enable this new architecture. GOS is the first demonstration that a Grid-enabled data communication protocol can accelerate tenfold distributed applications including OpenOffice, MySQL/IBM DB2, Firefox, MPlayer, and Google Earth, etc.

**BIO:**

Prof. Frank Wang is a Professor and Chair in e-Science and Grid Computing, Director of Centre for Grid Computing within the context of Cambridge-Cranfield High Performance Computing Facility (CCHPCF). Prof Wang's appointment is seen as crucial to the initiative of the CCHPCF, which is a collaborative research facility in the Universities of Cambridge and Cranfield. Prof. Wang has a publication record including a book titled "Encyclopedia of Grid Computing", 59 journal papers (12 IEEE Transactions, 5 JAP, 1 APL, 2 ACM publications, etc.) and 36 conference papers. His latest journal paper is *Grid-Oriented Storage: A Single-Image, Cross-Domain, High-Bandwidth Architecture* on IEEE Transactions on Computers, April 2007. Prof. Wang is on the Editorial Board of 4 international journals. He serves the High End Computing Panel for Science Foundation Ireland (SFI). He has been elected as the Chairman (UK & Republic of Ireland Chapter) of the IEEE Computer Society from January 2005.

# References


- [1] Network File System, Sun Microsystems, www.sun.com, 2007
- [2] CIFS: A Common Internet File System, Microsoft, www.microsoft.com, 2007
- [3] Andrew S. Tanenbaum, Computer Networks, 4th Edition, ISBN-10: 0-13-066102-3, Prentice Hall, 2002
- [4] David X. Wei, Cheng Jin, Steven H. Low and Sanjay Hegde, Fast TCP: motivation, architecture, algorithms, performance, IEEE/ACM Trans. on Networking, to appear in 2007.
- [5] Breaking Moore's law, A brief history of the Grid, GridCafe, gridcafe.web.cern.ch, 2006
- [6] GlobusWORLD, www.globusworld.com/program/program.php, 2006
- [7] Bill Allcock, Joe Bester, John Bresnahan, Ann L. Chervenak, Ian Foster, Carl Kesselman, Sam Meder, Veronika Nefedova1, Darcy Quesne, Steven Tuecke, Secure, Efficient Data Transport and Replica Management for High-Performance Data-Intensive Computing, 2000
- [8] Balázs Kónya, GridFTP tests over the NorduGrid Resources, 2nd NorduGridWorkshop Oslo, 2001
- [9] G. Gibson (Panasas Inc. & Carnegie Mellon), B. Welch (Panasas Inc.), G. Goodson, P. Corbett (Network Appliance Inc.), Internet-Draft, Parallel NFS Requirements and Design Considerations, October 18, 2004
- [10] The Panasas ActiveScale File System, www.panasas.com/panfs.html, 2006
- [11] P. H. Carns, W. B. Ligon III, R. B. Ross, and R. Thakur, ``PVFS: A Parallel File System For Linux Clusters'', Proceedings of the 4th Annual Linux Showcase and Conference, Atlanta, GA, 2000
- [12] IBM General Parallel File System, www-03.ibm.com/, 2006
- [13] E. Cohen, H. Kaplan and J. Oldham, "Managing TCP Connections under Persistent HTTP", Proceedings of the Eighth International World Wide Web Conference, Toronto, Canada, May 1999
- [14] C. Baru, R. Moore, A. Rajasekar, and M. Wan, “The SDSC Storage Resource Broker.” In Procs. Of CASCON'98, Toronto, Canada, 1998
- [15] S. Floyd. “Congestion Control Principles”, RFC2914.
- [16] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, Modeling TCP throughput: a simple model and its empirical validation. ACMSIGCOMM, 1998.
- [17] Web100 project, http://www.web100.org, 2007
- [18] Nettimer Project, Stanford University, mosquitonet.stanford.edu/~laik/projects/nettimer/, 2005
- [19] Jie Chen, Walt Akers, Ying Chen and William Watson III, Java Parallel Secure Stream for Grid Computing, http://www.ihep.ac.cn/~chep01/abstract/10-008.htm, 2005
- [20] Frank Wang, Sining Wu, Na Helian, Andy Parker, Yike Guo, Yuhui Deng, Vineet Khare, Grid-oriented Storage: A Single-Image, Cross-Domain, High-Bandwidth Architecture, IEEE Transaction on Computers, ISSN: 0018-9340, accepted, Vol.56, No.4, 2007.
- [21] Grid Security Infrastructure (GSI), www.globus.org/security/, 2007
- [22] Cambridge-Cranfield High Performance Computing Facility (CCHPCF), http://www.hpcf.cam.ac.uk/
- [23] OpenOffice, www.openoffice.org, 2007
- [24] Glenn Colaco (Sun Microsystems) and Darrell Suggs (Network Appliance), Database Performance with NAS: Optimizing Oracle on NFS, Technical Whitepaper, 2004
- [25] Robin Schumacher, MySQL Developer Zone, MySQL AB, http://www.mysql.com/, 2004
- [26] MySQL AB, MySQL Internals Manual, (revision: 472), 2005
- [27] International sequence databases exceed 100 gigabases, www.ncbi.nlm.nih.gov/Genbank/, 2007
- [28] MPlayer - The Movie Player for Linux, www.mplayerhq.hu, 2007
- [29] Mozilla Firefox web browser, www.mozilla.com, 2007
- [30] Explore Google Earth, earth.google.com, 2007
- [31] Cheng Jin, David X. Wei, Steven H. Low, G. Buhrmaster, J. Bunn, D. H. Choe, R. L. A. Cottrell, J. C. Doyle, W. Feng, O. Martin, H. Newman, F. Paganini, S. Ravot, S. Singh, Fast TCP: from theory to experiments, IEEE Network, 19(1):4-11, January/February 2005.
- [32] Tom Kelly, Scalable TCP: Improving Performance in Highspeed Wide Area Networks. Computer Communication Review 32(2), April 2003
- [33] Multi-core (computing), http://en.wikipedia.org/wiki, 2007

# QUESTIONS ?