eXludus Technologies, Inc.

RepliCator:

Optimization Software Suite

for Clusters & Grids

February 2007

Our Objective :
Cluster and Grid Optimization

- Through Intelligent Data Transfer Management

- For enhanced ROI and ease-of-use

- By providing a full set of tools for:

  - Automated coordination of Data Movement

  - Simplified movement of Data between multiple sources

  - User transparency for local and remote Data transfers

  - Enhanced performance of Data transfers

  - Support of Data-driven scheduling and processing

Grid & Cluster :
Typical Landscape

How to manage complexity?

Remote Data

Remote Queue Host

Queue Host

Cluster #1

Cluster #2

Scratch Storage

Central Data Server
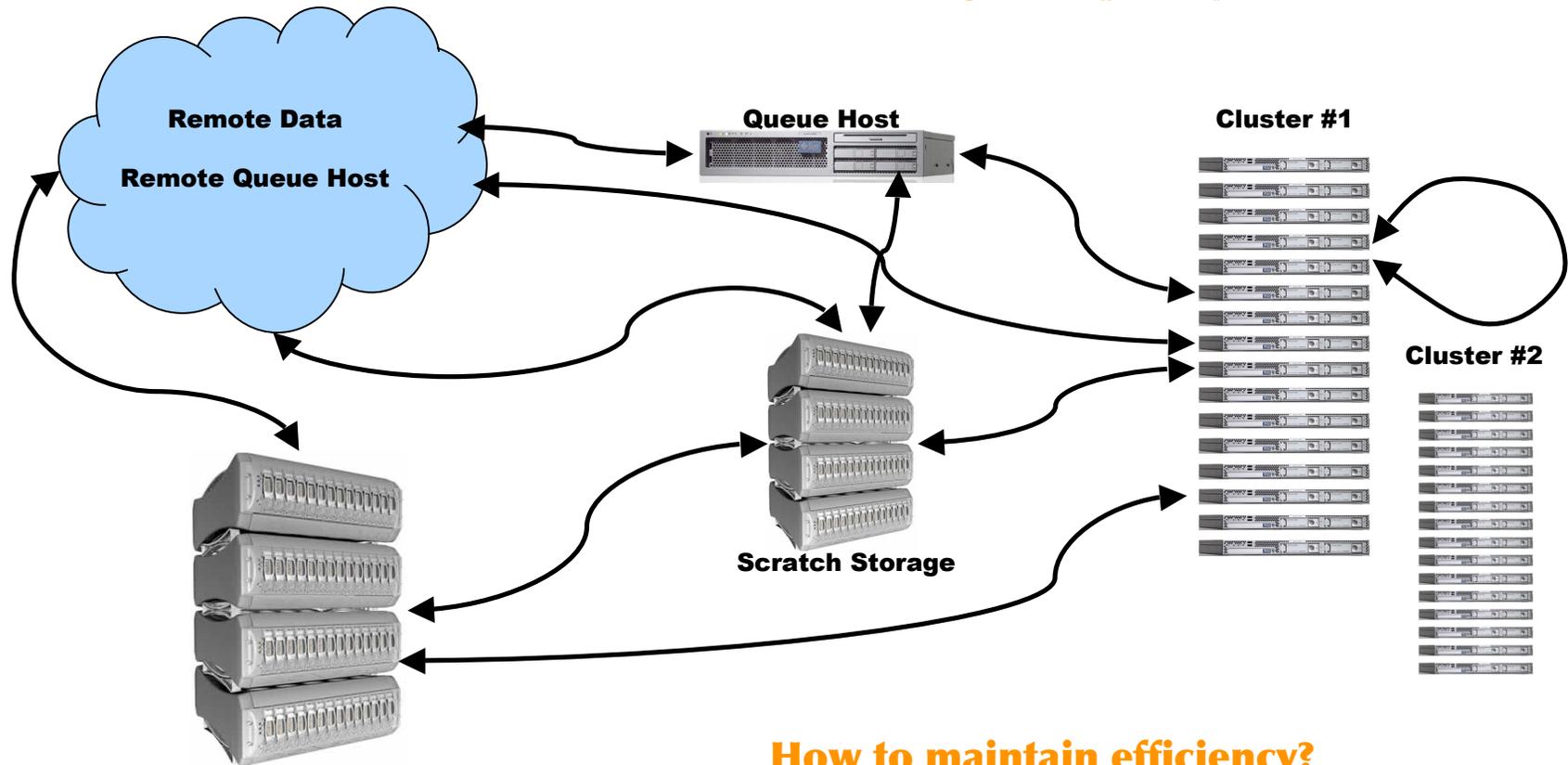
How to maintain efficiency?

RepliCator's Impact :
Data Center Pain Killer!
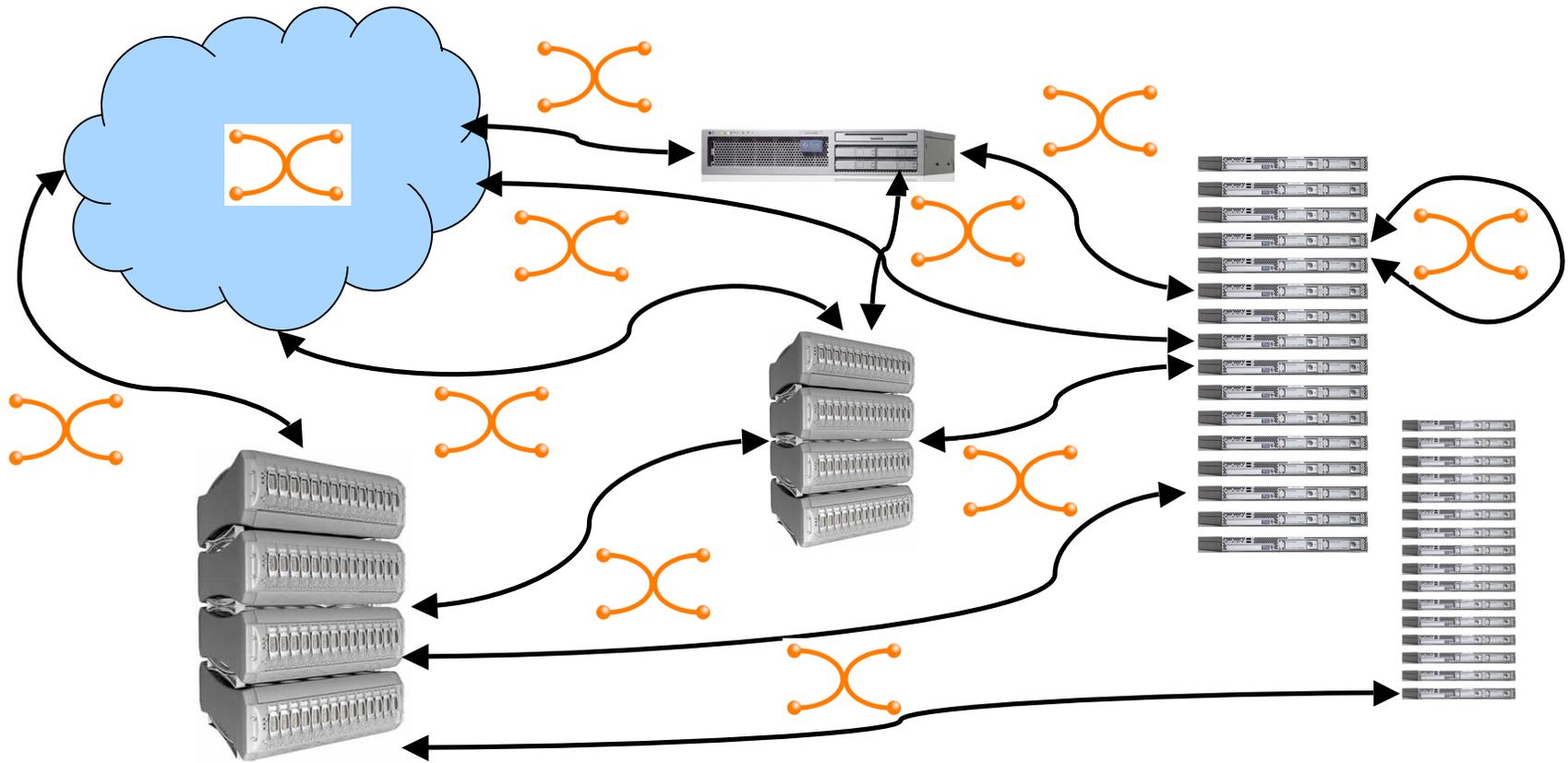
What RepliCator Does :
Cluster and Grid Optimization

- ✕ Data Transfer

- ✕ Data Activated Processing

- ✕ Data Aware Scheduling

- ✕ Data Synchronization

- ✕ Job Submission Control

- ✕ Process Scheduling Control

- ✕ Information Feedback

We automate and optimize Cluster and Grid data transfers without expensive add-ons

VirtualWare

Processor

Gridware

Network
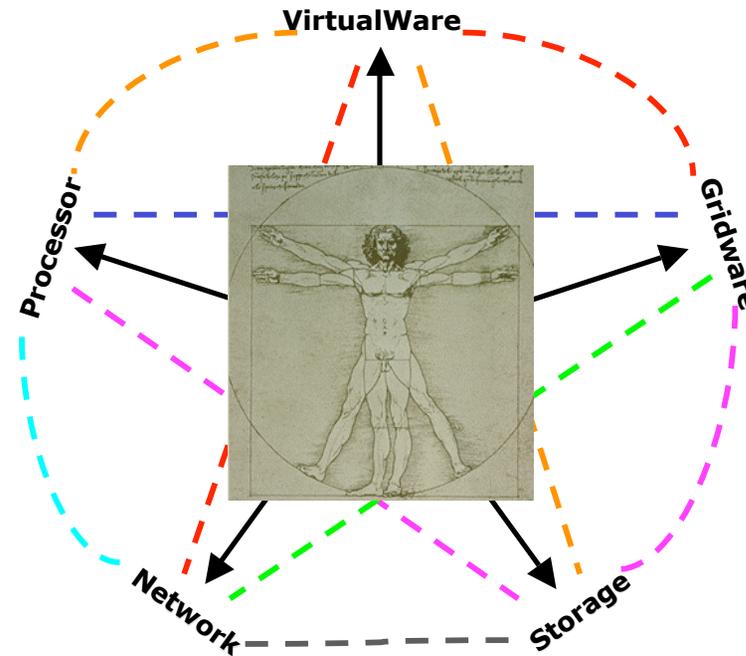
Storage

RepliCator UI & API :
Data Transfer Optimizer Module

- Seamlessly transfer data from remote locations to central storage or scratch storage facilities

- Seamlessly transfer data from central storage to scratch storage
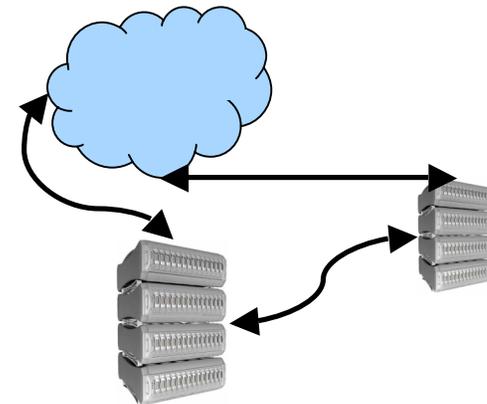
```
replic-qsub -F nt job.sh
```

RepliCator MLP :
Job Dispatch Optimizer Module

- Control job dispatch based on Data locality (Data-Activated Processing)

- Control job dispatch based on node state

- Control job dispatch based on user determined conditions

- Transparent to users

- Interacts with workload managers

```
DATA {TRANSFER file … ; }

EXECUTE {DO program …  ; }

CLEANUP {PURGEALL; }
```
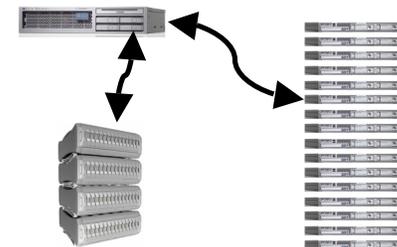
RepliCator ART :
Asynchronous Result Transfer Optimizer

- Automatic transfer of data back to

  - scratch storage

  - central storage

  - remote location storage

  - peer node local disk

- On-the-fly transfer of data while

  processing continues (eg. checkpointing)

- External triggering of data transfers

Processing Efficiency

RepliCator

NFS

% Efficiency
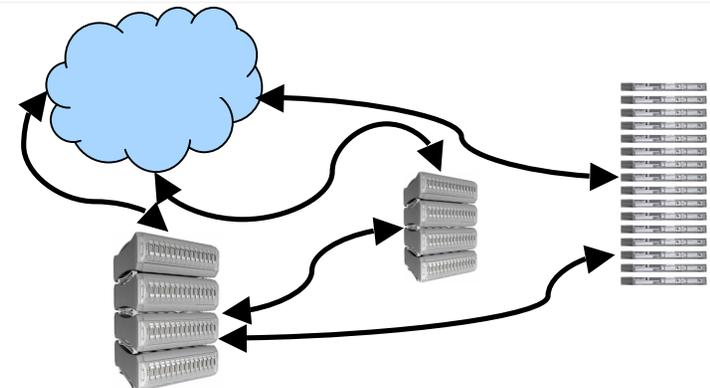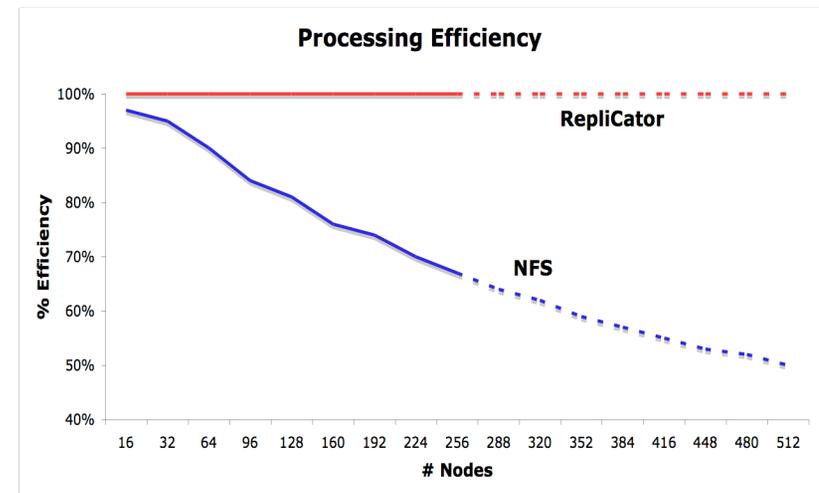
Nodes

RepliCator : Parallel File Optimizer

- Cache data on multiple compute nodes simultaneously
  - Scratch Storage
  - Central Storage
  - Remote Locations

Aggregate Network Bandwidth

RepliCator

MBy/s

Panasas

NFS

nodes

9
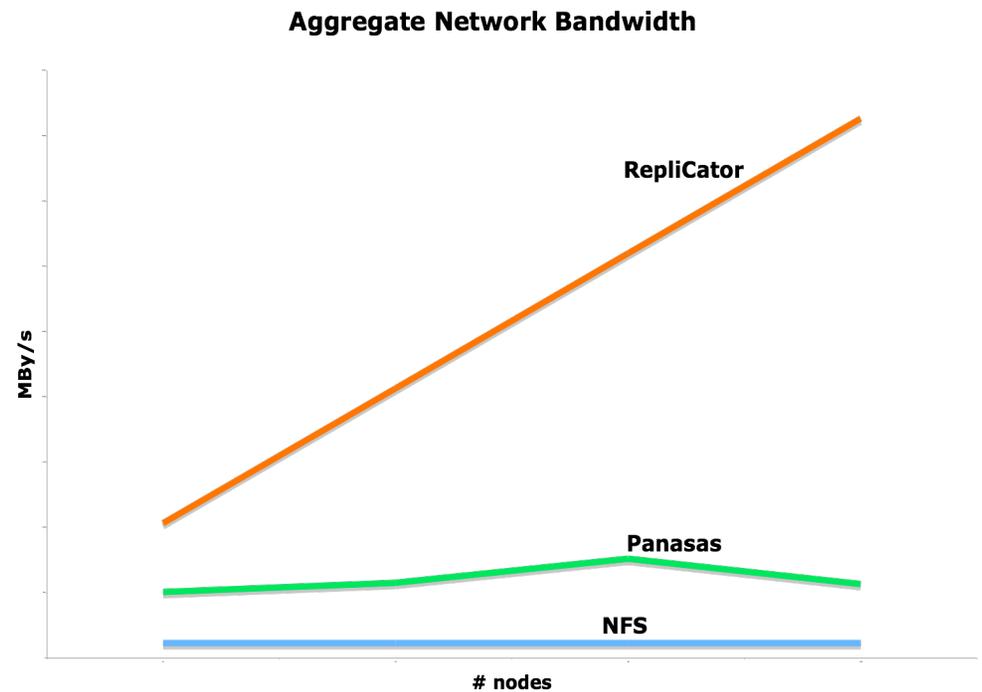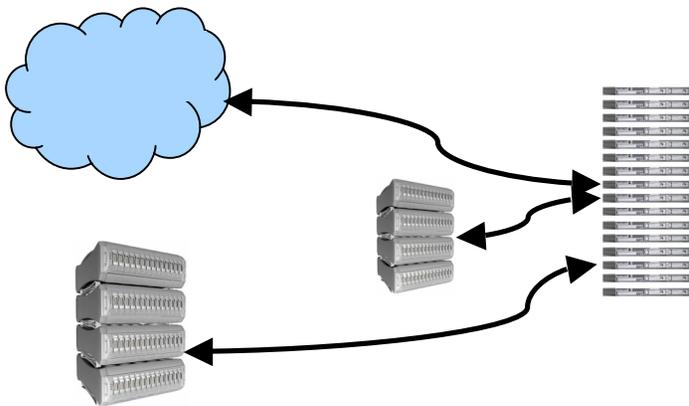
RepliCator : SchedOptimizer
Process Scheduling Optimization

⚬ Automatic optimization of node
processing capacity via interaction
with workload scheduler

Efficiency Gain

Speedup

Initial   Stage 1   Stage 2   Stage 3   Stage 4

Optimization Stage

SchedOptimizer performance gain is incremental to parallel file serving and ART performance gains.
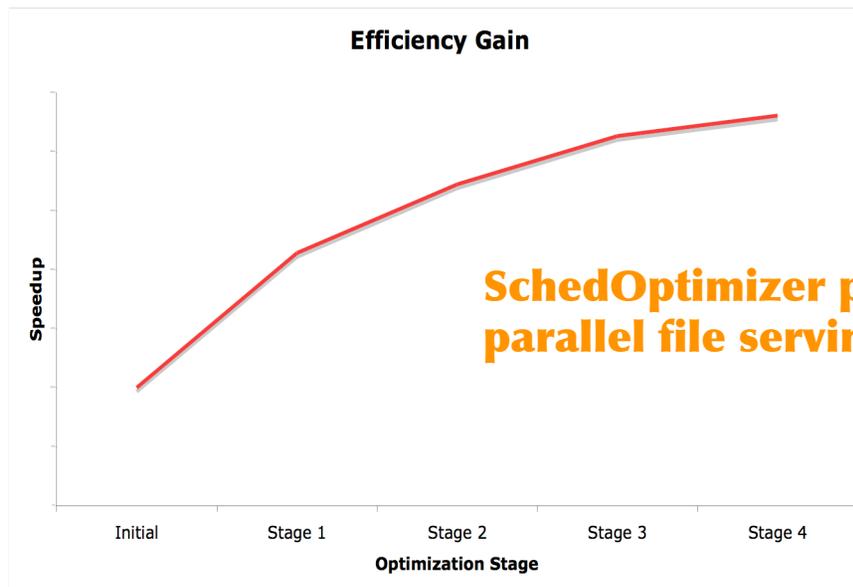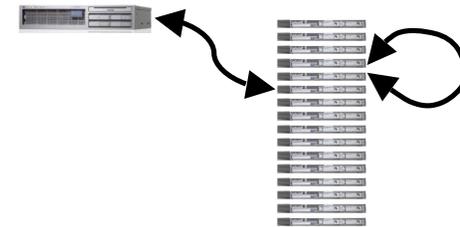
Replicator Scalability
BLAST-NT Parallel File Serving Benchmark (out of memory)

In core / out of core,

Small / large cluster,

Fast / slow network,

NFS / Cluster FS,

Single / multi core,

RepliCator *is* faster.

Way faster.

CPU Efficiency

%CPU Used vs # nodes

- RepliCator — ~35% flat
- NFS
- Panasas

X-axis: 16, 32, 64, 96, 128 (# nodes)
Y-axis: 0% to 40% (%CPU Used)
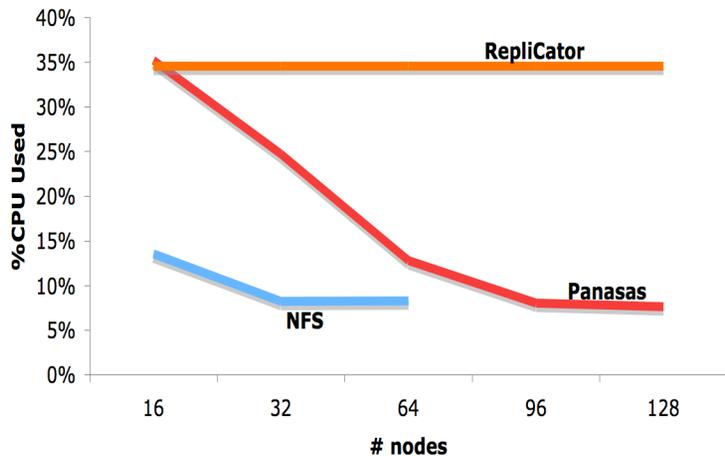
2 Single-Core AMD cpu nodes
3.7 GB memory nodes
550 MB/s Panasas

BLASTN: 64 Nodes

- RepliCator: **2804 s**
- Panasas: **6795 s**
- NFS: **26743 s**

Legend: User, System, WaitIO
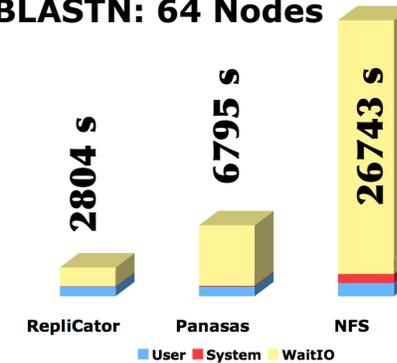
BLASTN: 128 Nodes

- RepliCator: **1404 s**
- Panasas: **6818 s**
- NFS

Legend: User, System, WaitIO

11
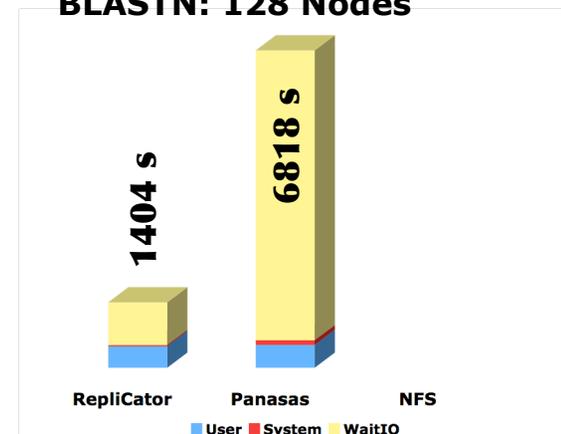
Replicator Scalability
BLAST-NT Parallel File Serving Benchmark (in memory)

39 Nodes
100% clean cache

Lustre: 118 s, RepliCator: 35 s
USER ■ SYS ■ WAITIO

RepliCator *is* **faster.**

Way faster.

39 Nodes
Initial DB loading

Lustre: 788 s, RepliCator: 37 s
USER ■ SYS ■ WAITIO
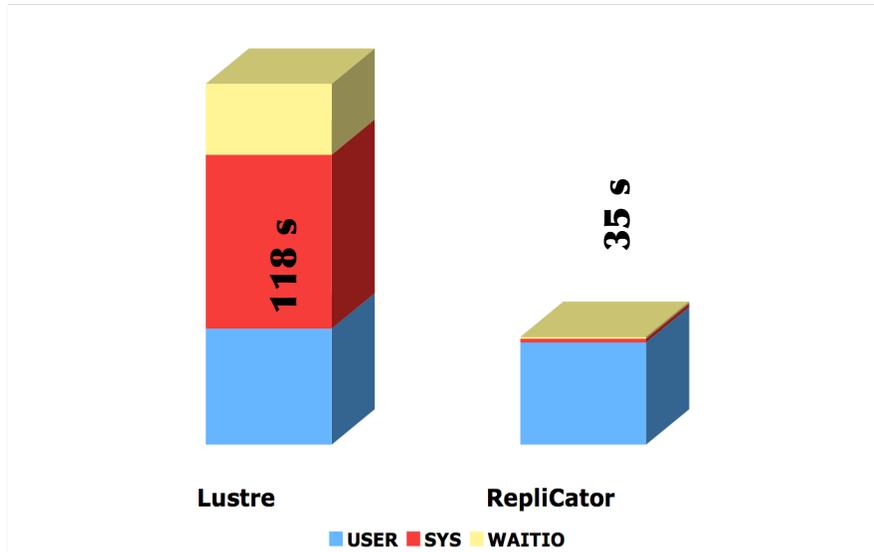
2 Dual-Core Intel cpu nodes
8 GB memory nodes
1.2 GB/s Lustre
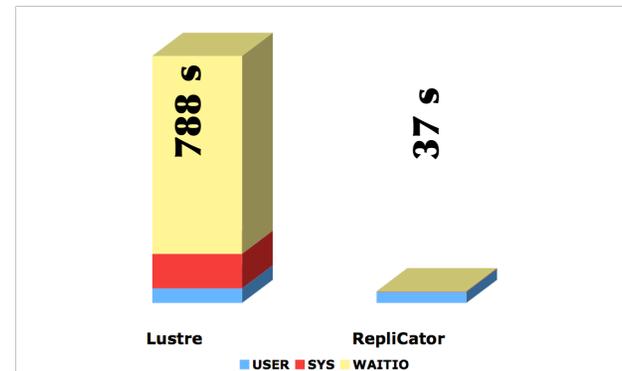
RepliCator Suite Interoperability

RepliCator Ports :	SGE	PBS	LSF	Torque
Linux 32 bits	Now	Now	Q1	Now
Linux 64 bits	Now	Now	Q1	Now
Solaris 64 bits	Now	Now	Q1	Now
Mac Os X	Now	Now	Q2	Q2
Windows 32 bits	Now	Now	Q2	Q1

Nearly any mix of OS, Chipset, WLM, network, file system
(NFS, Lustre, Panasas, Ibrix, …).

No modifications to your hardware, file system, networking infrastructure

13

RepliCator Optimization Capabilities

✕ **Automation :** automate transfer to scratch, submit job, cleanup and retrieve results

✕ **FT - Resilient Transfer :** able to survive and recover transfer interruptions (ie RCP, FTP can't)

✕ **NetDiag :** network diagnostic tools for storm, suppression, duplication, BW, etc. testing

✕ **Parallel File Serving Optimizer :** 1-to-N nodes stage in

✕ **ART Optimizer :** stage out (NFS-CFS, peer, remote; command line and API)

✕ **ART-trigger :** on-the-fly external trigger mechanism

✕ **Meta Language Processor :** event driven mechanism at nodes; WLM driver at server

✕ **FTP :** remote operation & interface to portals

✕ **VM Provisioning Optimizer :** enable asynchronous virtualware deployment in large clusters

✕ **SchedOptimizer :** multi-core / CPU RT scheduling optimization (job arrays)

Keeping Costs Down, Enhancing ROI

✂ Clusters are commodity, what about infrastructure?

✂ RepliCator Grid Optimizer boosts infrastructure ROI:

  ✂ Increased throughput

  ✂ Reduced job turnaround time

  ✂ Reduced costs of infrastructure

  ✂ Ease of use

  ✂ Transparent integration, management

  ✂ Straightforward licensing

Cluster Component Cost

Infrastructure

Commodity Processors

Cost

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%

1x
2x
4x
8x
16x

Size
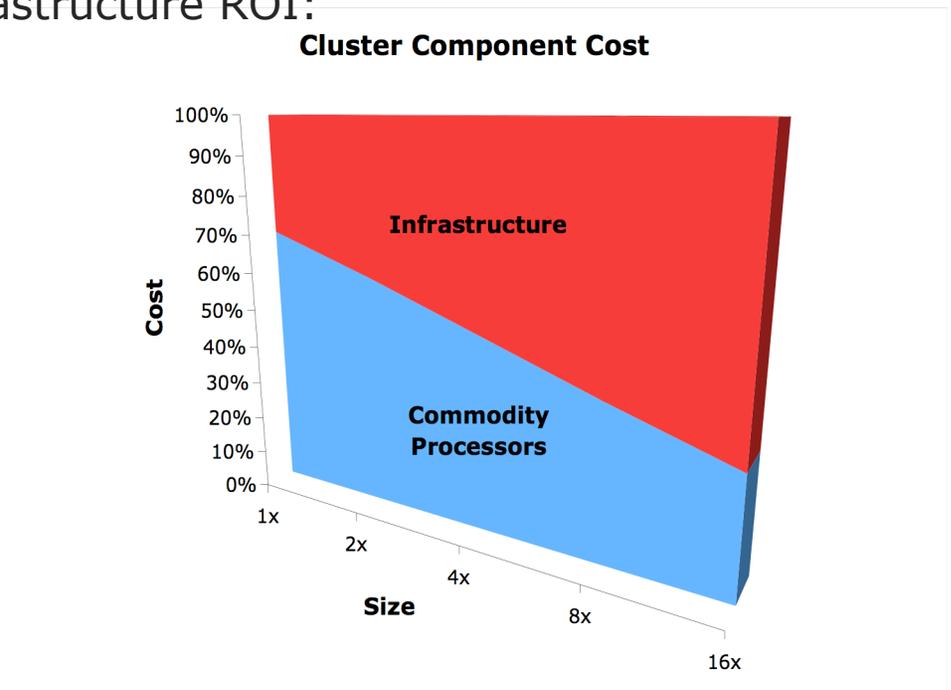
Testimonials

Bielefeld University/CeBiTec

"Setting up the RepliCator solution was really easy. No kernel-tweaking or similar changes were involved. eXludus was very helpful, and we set up a first running replication in a matter of hours, and a good, working software-integration in less than a week. … the effective network bandwidth for the replicated files is now almost independent of the number of nodes. … there is really lots of headroom for future cluster extensions."

> - Ralf Nolte, System admin., CeBiTec BRF

York University

"The initial experiments conducted by the University of York and eXludus showed that *even for a cluster with as few as 10 nodes, RepliCator greatly improved system utilization* by accelerating data flow 4.4 times in comparison with the widely used NFS file system,"

> - Aaron Turner, White Rose Grid manager at the University of York.

Cambridge University

"For I/O limited jobs this can dramatically increase throughput, we have seen *throughput increases of over 100% on certain codes*"

> - Paul Calleja, Director HPCF

Top Ten Bio-It Trends for 2007
Bio-It World, Dec06 - Jan07

Trend #2. "Data-aware" Computing :

Companies such as Exludus can layer file-and-data locational awareness onto

common distributed resources managements systems such as Platform LSF and

Sun Grid Engine. The result is that scientific workflows automatically migrate

towards systems where the required datasets already reside or can be quickly

obtained. For environments with I/O heavy scientific computing requirements the

throughput ***performance gains can be impressive***.

> **- The BioTeam**
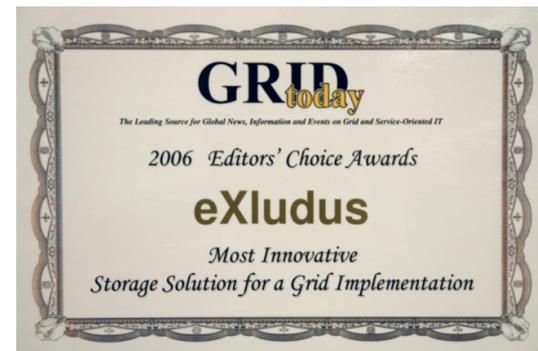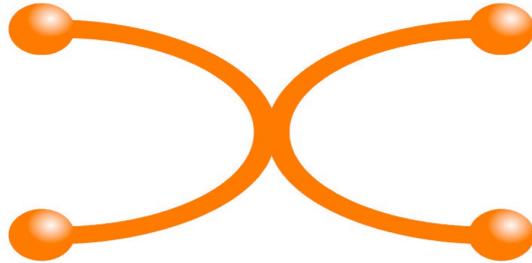
Offices:

HQ: Montréal, Québec, Canada
  - Benoît Marchand, development & support team

Americas Sales: Cupertino, California, USA
  - Stephen Perrenod

EMEA Sales: Kent, United Kingdom
  - Philip Bull

first.last@exludus.com
info@exludus.com
http://www.exludus.com

The Big Picture: Grid Optimizer

- RepliCator adds increased coherency, enhanced control and accelerated performance to the Data Center
  - Helps to implement "Next Generation Data Center"
  - Allows realization of Grid 2.0 practically, including support of SLAs
    - (Grid 2.0 = Grid 1.0 + storage, network, data management)
- RepliCator supports:
  - Multiple levels of Data hierarchy
  - Access to Data at remote sites
  - Small, medium or large clusters
  - Scaling in an economical fashion