Center for Integrated Turbulence Simulations

Building and Running Lots of Large Rocks Clusters

Steve Jones
Stanford University

Supported Groups

School of Engineering

Institute for Computational and Mathematical Engineering

- Flow Physics and Computation (FPC)
- Aeronautics and Astronautics
- Chemical Engineering
- Center for Turbulence Research (CTR)
- Center for Integrated Turbulence Simulations (CITS)
- Thermo Sciences Division (TSD)

Funding

- Sponsored Research (AFOSR/ASC/DARPA/DURIP/ASC)

Active collaborations with the Labs

Buoyancy driven instabilities/mixing - CDP for modeling plumes
(Stanford/SNL)

LES Technology - Complex Vehicle Aerodynamics using CDP
(Stanford/LLNL)

Tsunami modeling - CDP for Canary Islands Tsunami Scenarios
(Stanford/LANL)

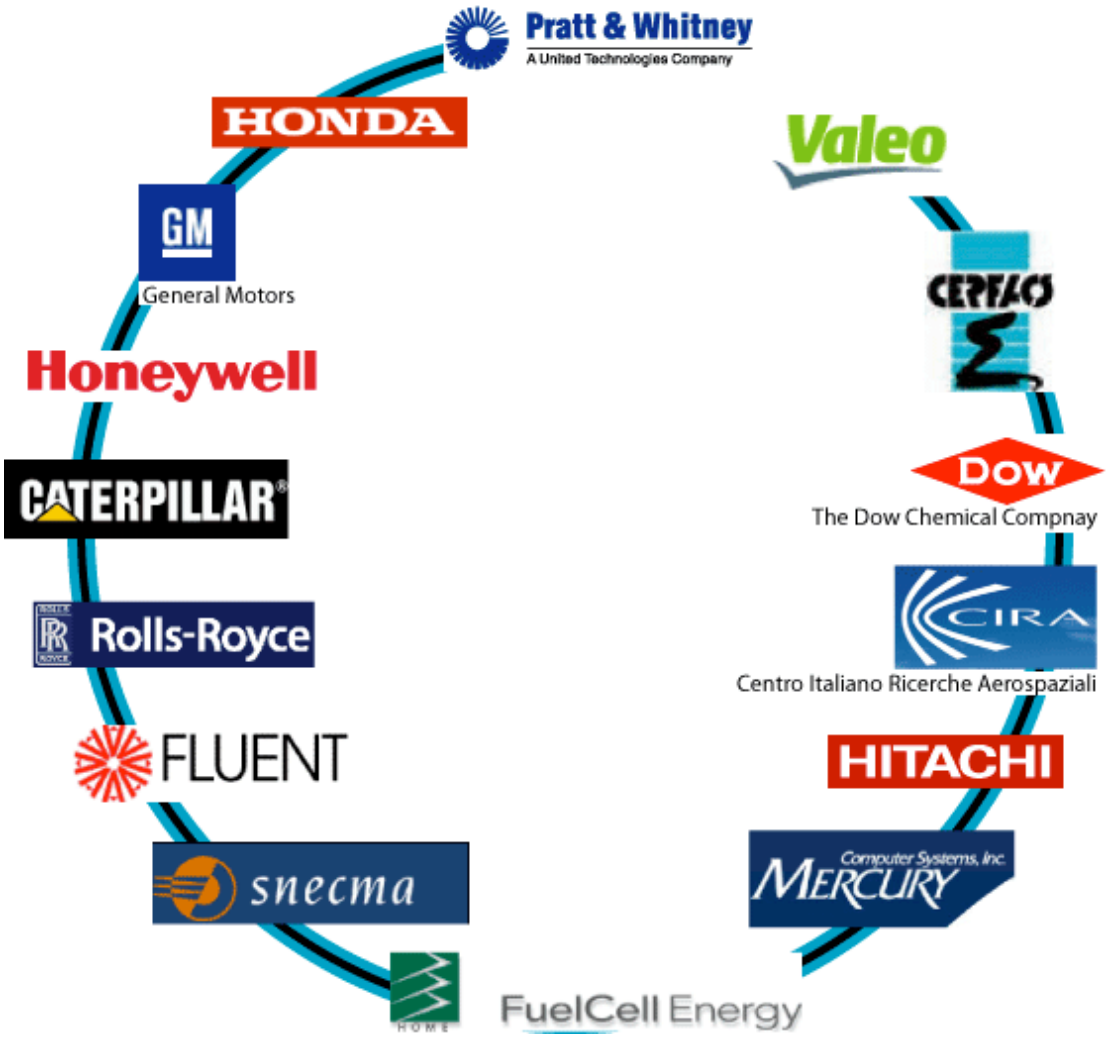
Parallel I/O & Large-Scale Data Visualization - UDM integrated in CDP
(Stanford/LANL)

Parallel Global Solvers - HyPre Library integrated in CDP
(Stanford/LLNL)

Parallel Grid Generation - Cubit and related libraries
(Stanford/SNL)

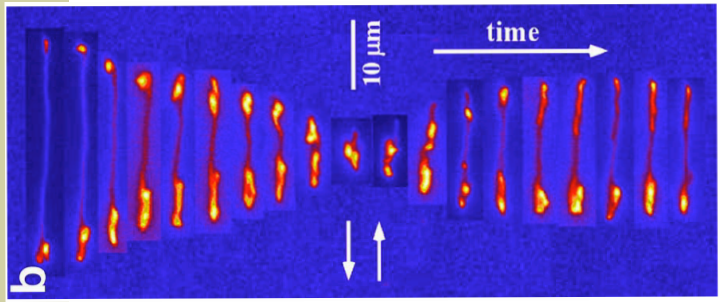
Merrimac - Streaming Supercomputer Prototype
(Stanford/LLNL/LBNL/NASA)

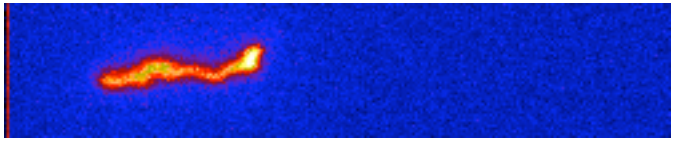
Affiliates Program

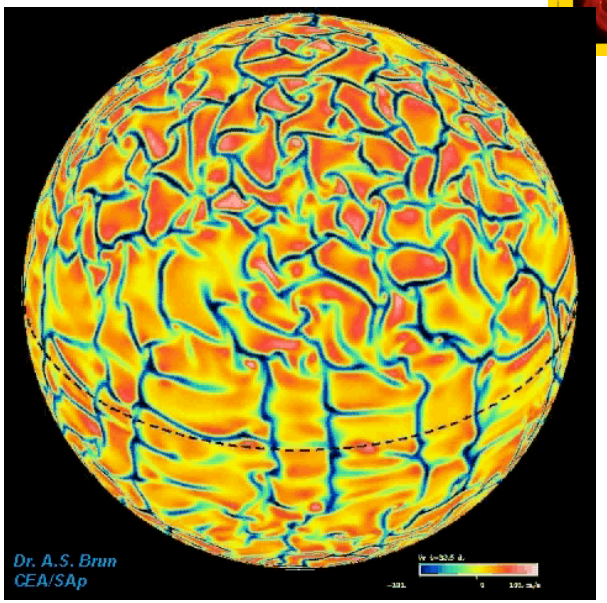


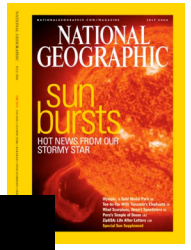
The Research

MOLECULES TO PLANETS !

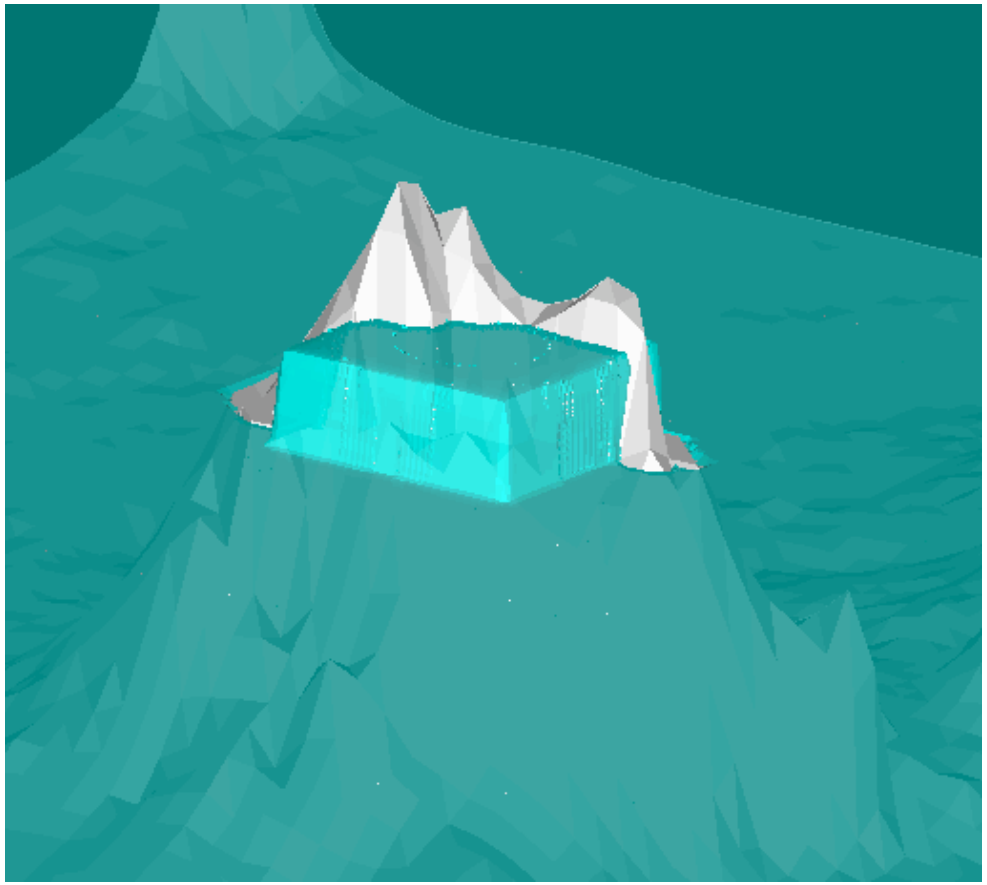




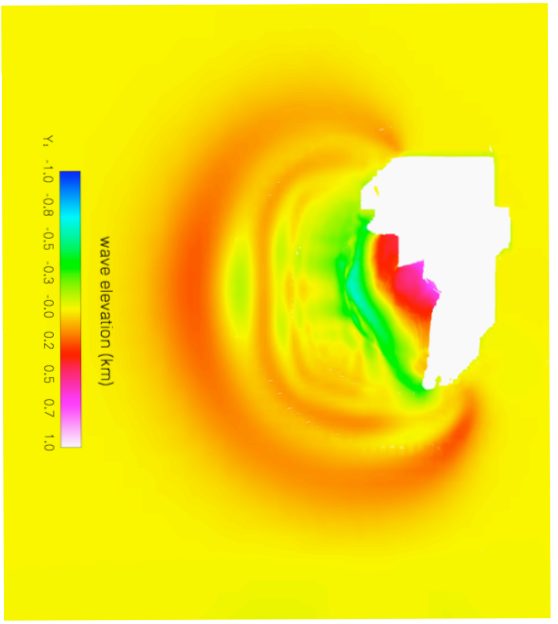




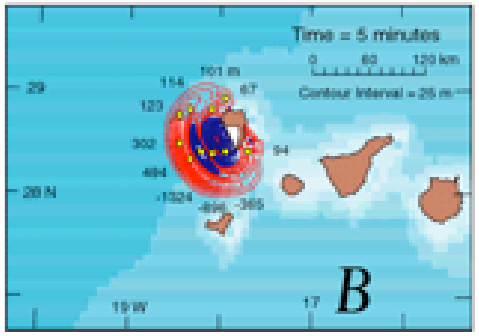
Tsunami Modeling



Preliminary calculations



Preliminary calculations



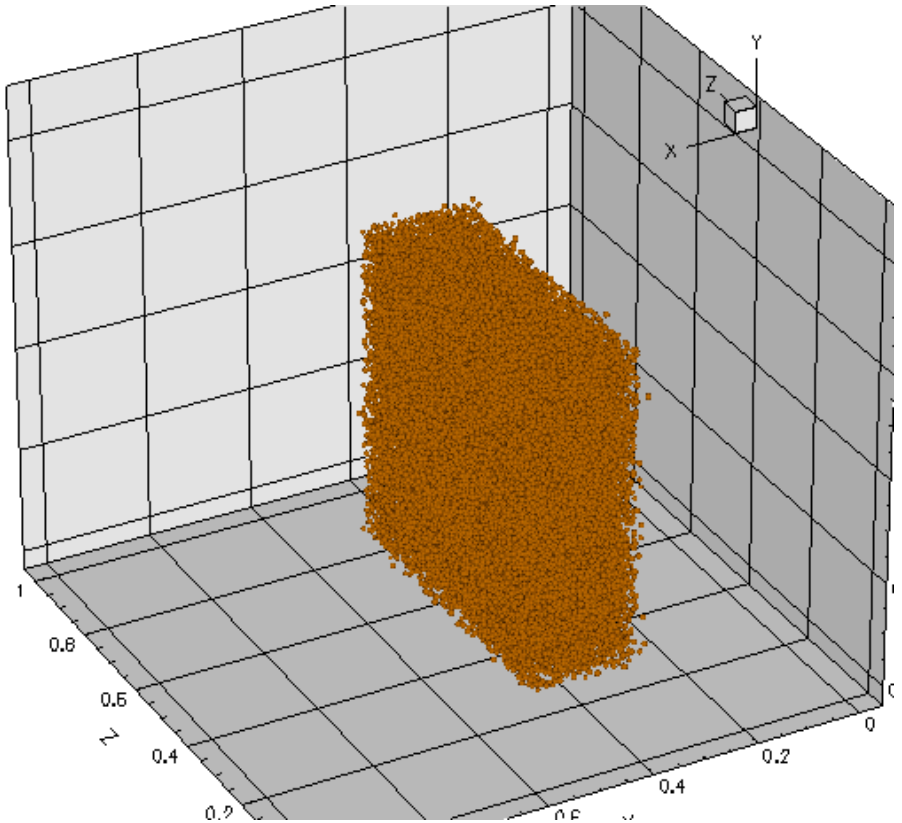
Ward & Day, Geophysical J. (2001)

Landslide Modeling

Extends existing Lagrangian particle-tracking capability in CDP

Collision model based on the distinct element method*

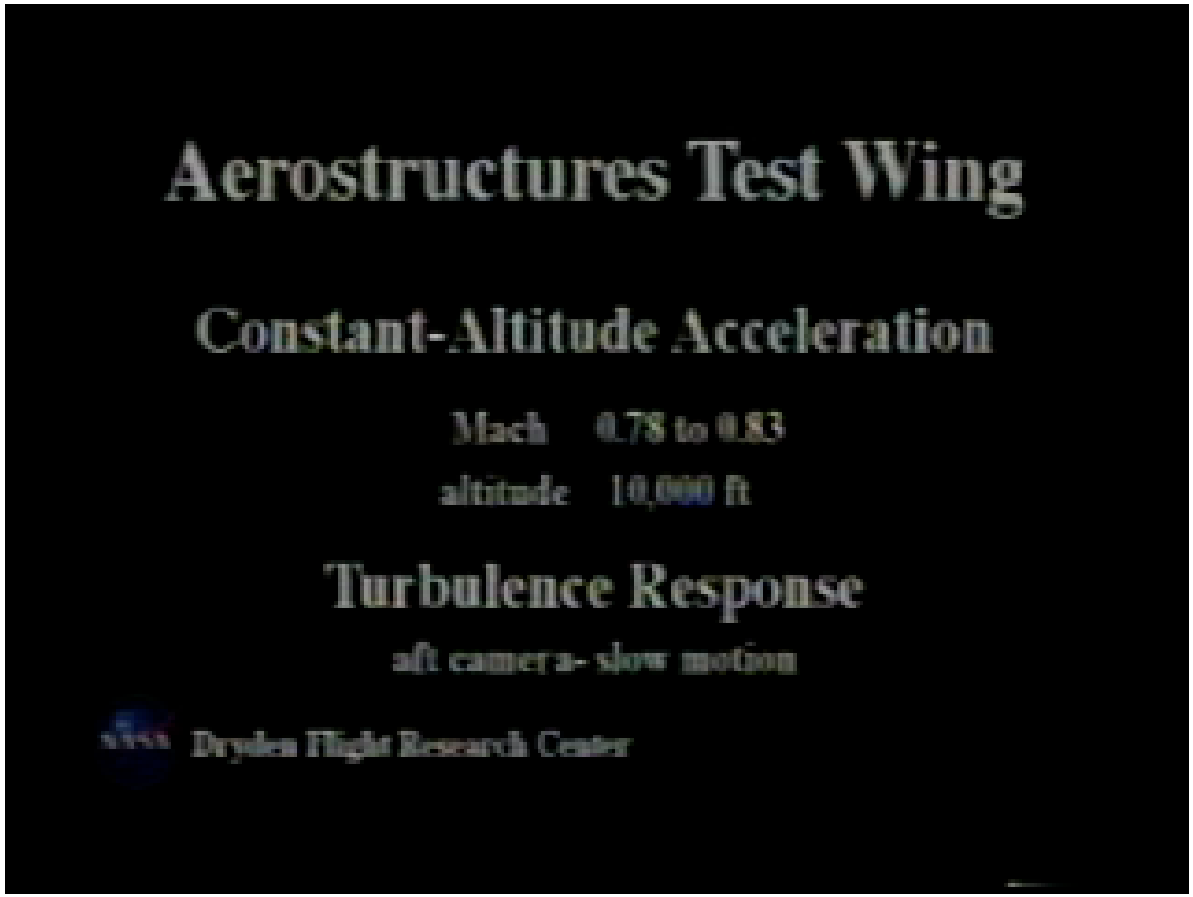
Originally developed for the analysis of rock mechanics problems



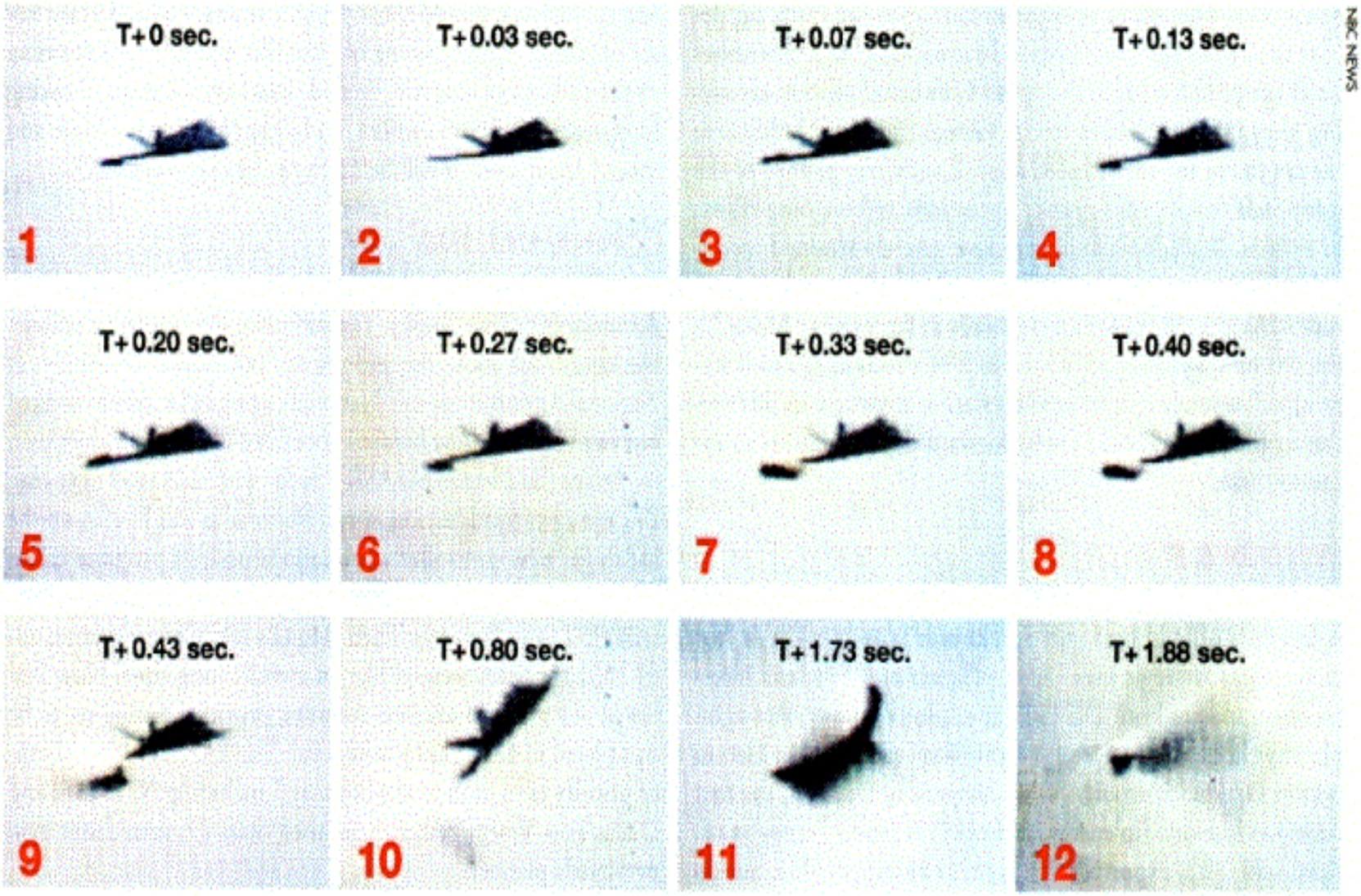
* Cundall P.A., Strack O.D.L., A discrete numerical model for granular assemblies, Géotechnique 29, No 1, pp. 47-65.

"Some fear flutter because they don't understand it, and some fear it because they do."

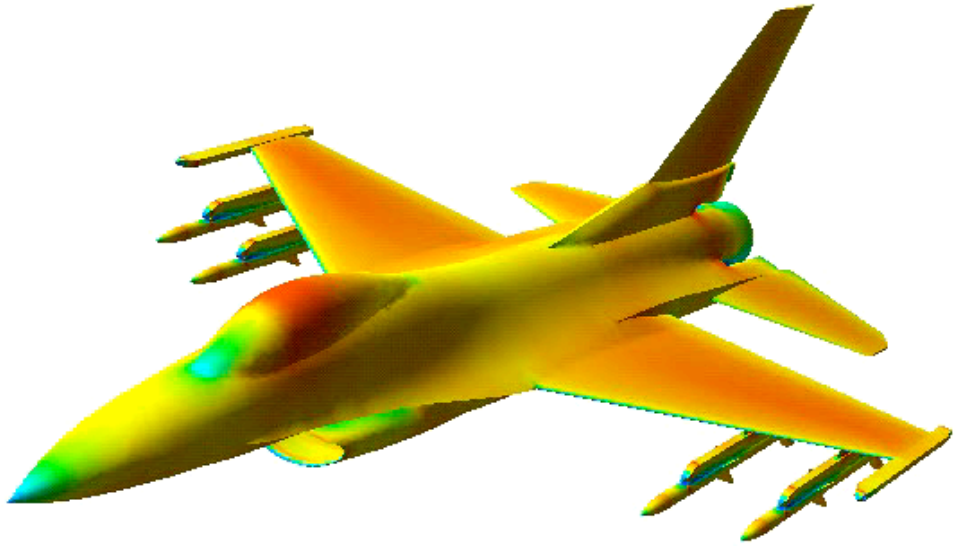
-von Karman-



9/12/97


T+0 sec.
1
T+0.03 sec.
2
T+0.07 sec.
3
T+0.13 sec.
4
NBC NEWS
T+0.20 sec.
5
T+0.27 sec.
6
T+0.33 sec.
7
T+0.40 sec.
8
T+0.43 sec.
9
T+0.80 sec.
10
T+1.73 sec.
11
T+1.88 sec.
12

Limit Cycle Oscillation

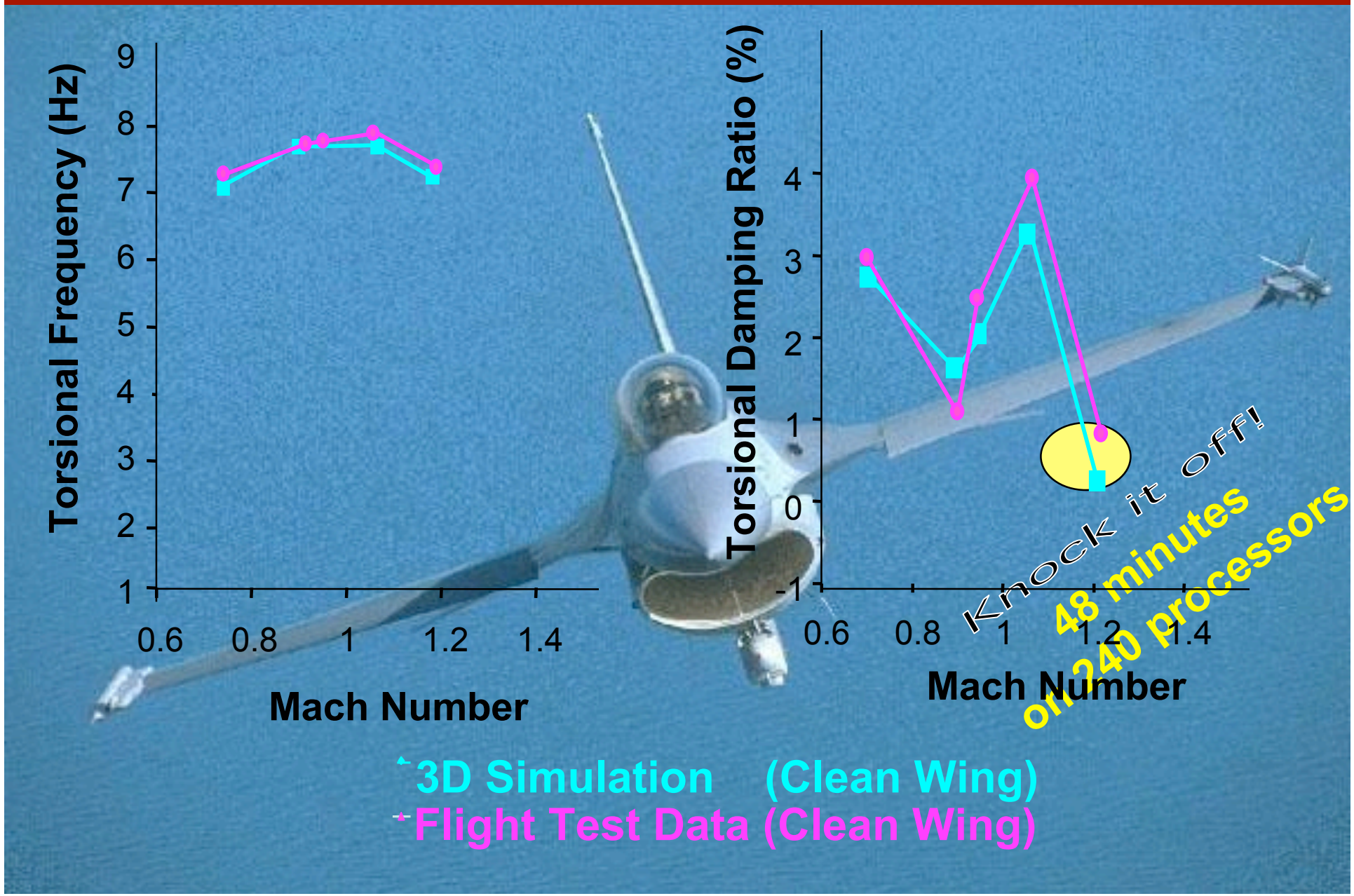


Fz_stores
Fz_clean

Time : 0.000000

Center for Integrated Turbulence Simulations

5G Turn



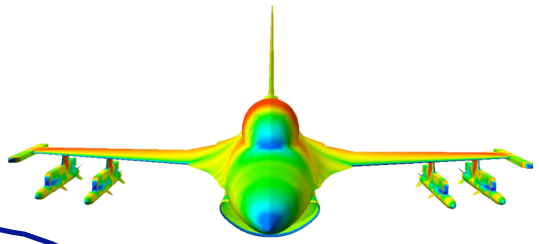
Databases?

Desert Storm (1991)

Iraq War (2003)

400,000 configurations to be flight tested

Potential



Damping Coefficient (%) -- 1st Torsion
4
3.5
3
2.5
2
1.5
1
0.5
0
0.6
0.8
1
1.2
Mach Number
Flight Test
FOM (1,170 s.)
TP-ROM (5 s.)

The Clusters

Iceberg

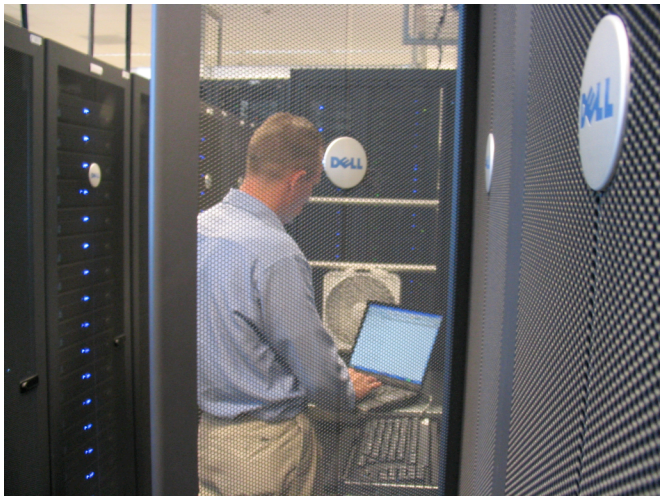
- 600 Processor Intel Xeon 2.8GHz
- Fast Ethernet

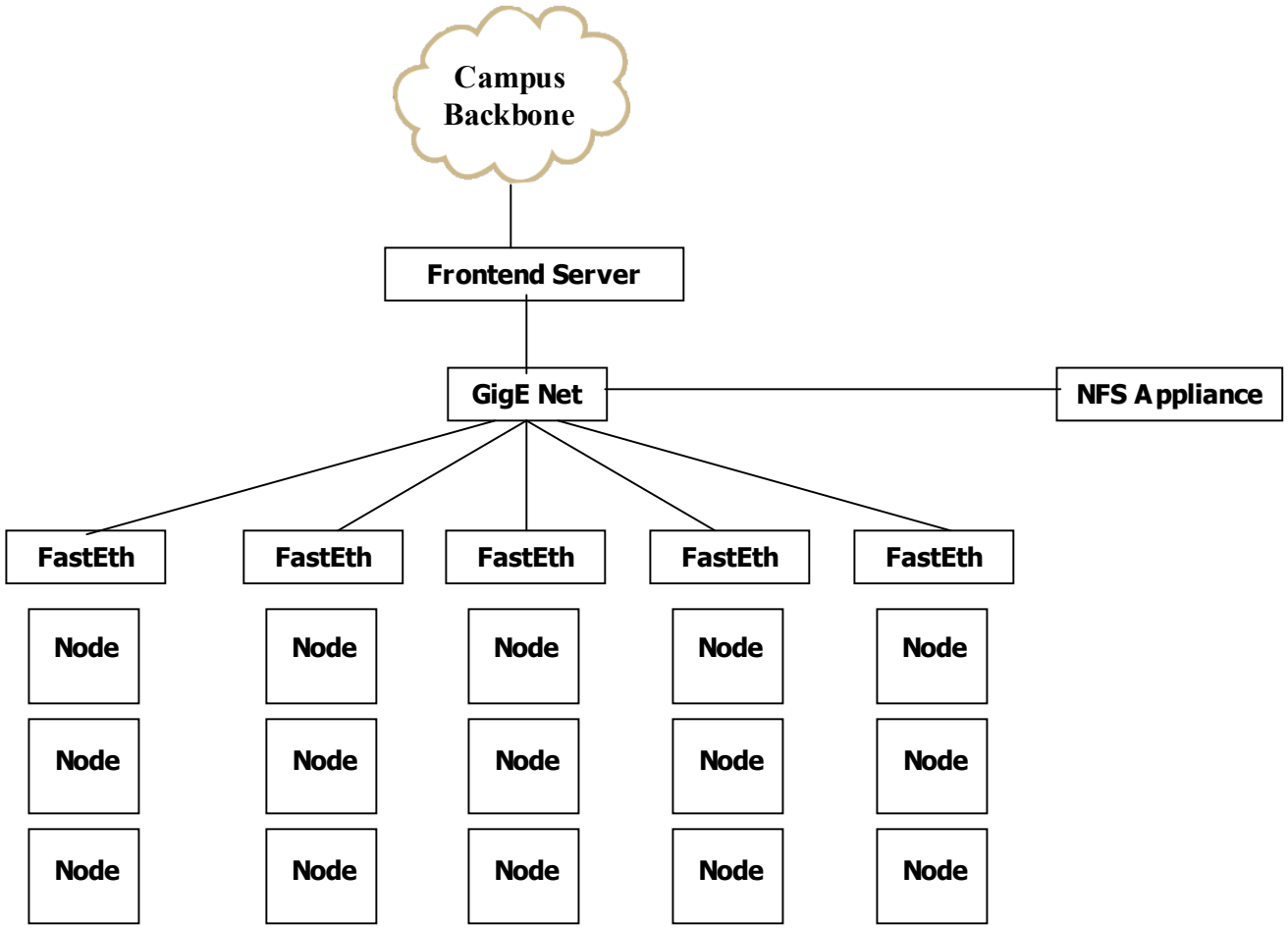
(Install Date 2002)

- 1TB Storage
- Physical installation - 1 week
- Rocks installation tuning - 1 week

Iceberg at Clark Center

One week to move and rebuild the cluster… then running jobs again




Campus Backbone
Frontend Server
GigE Net
NFS Appliance
FastEth
FastEth
FastEth
FastEth
FastEth
Node
Node
Node
Node
Node
Node
Node
Node
Node
Node
Node
Node
Node
Node
Node

Top 500 Supercomputer

TOP500® SUPERCOMPUTER SITES

Rank	Manufacturer Computer/Procs	Rmax Rpeak	Installation Site Country/Year
319	Dell PowerEdge 2650 Cluster P4 Xeon 2.8 GHz - Fast Ethernet/ 576	363.40 3225.60	Stanford University/Biomedical Computational Facility USA/2003
25	Dell PowerEdge 2650 Cluster P4 Xeon 2.4 GHz - Myrinet/ 600	2004.00 2880.00	University at Buffalo, SUNY, Center for Computational Res. USA/2002

eWEEK LABS

Supercomputers for the masses?

LABS ON-SITE AT STANFORD: PC CLUSTER SHOWS ENTERPRISE POTENTIAL

By John Taschek

TEN YEARS AGO, SUPERcomputers were multimillion-dollar systems usually used for massive projects, such as modeling Earth's climate or nuclear reactions. Today, they are called HPCCs, or high-performance computing clusters, and they are basically free when built from spare PCs. More important, they are quickly becoming suitable for mainstream enterprise computing.

HPCCs look completely different from traditional supercomputers: They are fan-cooled, not water-cooled, and they sit in racks and use off-the-shelf components. And while the inventor of supercomputers—Cray Research Inc.—may have cranked out only two or three computers a year a decade ago, companies including Dell Computer Corp., Red Hat Inc. and Microsoft Corp. are now building hundreds of postmodern supercomputers at a time.

The changes in supercomputing can be seen most clearly in academia, where the New Age supercomputers are commonly used.

eWEEK Labs recently visited Stanford University, in Stanford, Calif., which was setting up a 300-node cluster together the different sciences—including engineering, physics, medicine and biology—so researchers can better share resources, planning and data.

Stanford began building its cluster late last year, with help from Dell and Intel Corp. By last month, the University of Tennessee and the National Energy Research Scientific Computing Center (see excerpted list, right).

Ironically, just a few years ago, the Top 500 list comprised mostly SGI systems based on Cray technology. Now, there is just a sprinkling of Crays in the mix.

The fastest Cray on the list is at No. 39, clocking in at 1,166 gigaflops—nearly a thousand times faster than a Cray Y-MP circa 1988. Interestingly for the enterprise, the performance of the No. 39 Cray system, which is used by the government for unknown but probably defense-related

Steve Jones (left) and eWEEK Labs Director John Taschek running the supercomputer benchmark at Stanford. At left, the Stanford cluster.

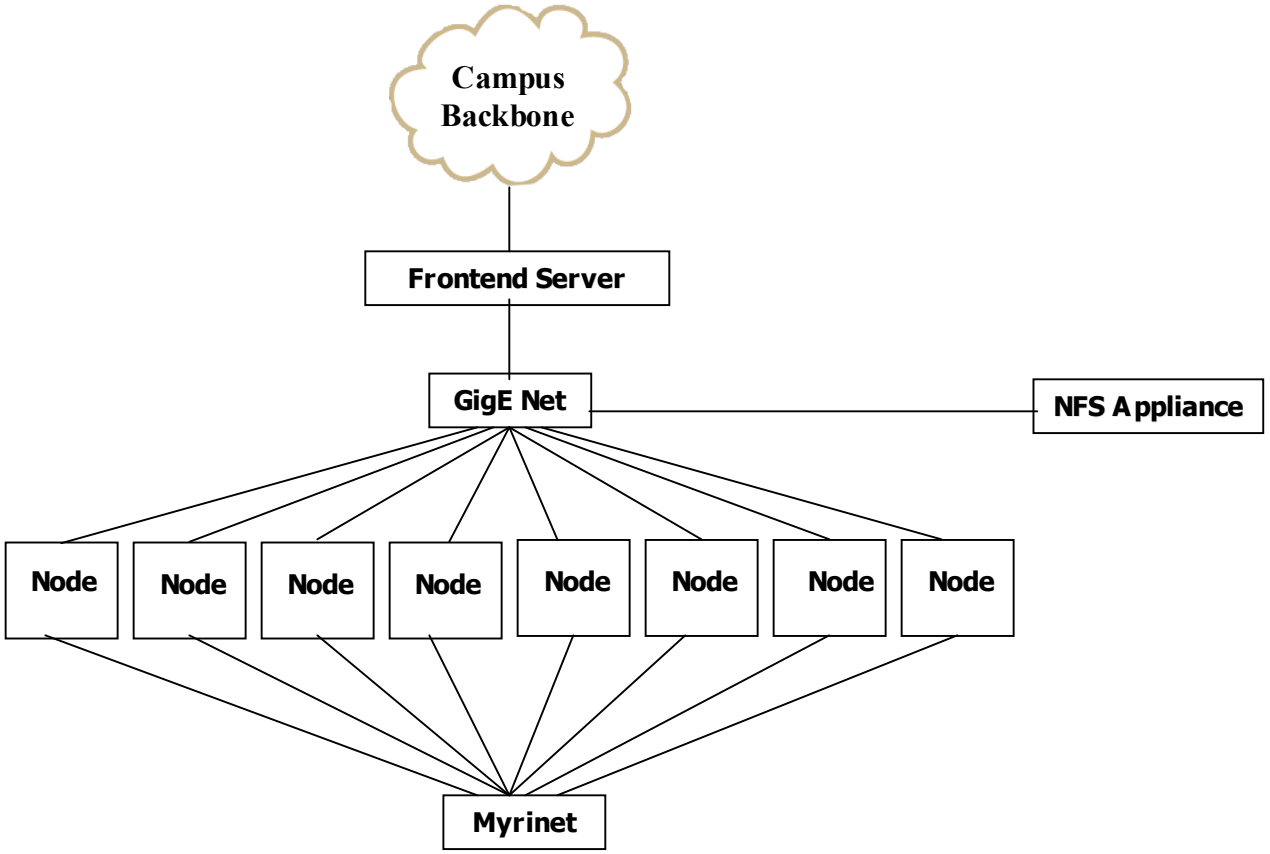
Stanford's original goal was to place in the first 70 of the Top 500. However, after the system was built, Steve Jones, architect of the Stanford clustering project, said the best he hoped for was a spot in the first 200 in the benchmark. (The numbers are still being crunched, but eWEEK estimates that the Stanford system will come in at about 170.)

Although low-cost computers can be used in a cluster, the network switching fabric has a significant impact on performance. Because of cost concerns, Jones was forced to use a 100BaseT (Fast Ethernet) network backbone instead of the far-faster Gigabit Ethernet fabric. "The switching fabric has a huge impact on our placement for the Top 500 list," said Jones. "Due to costs, we sacrificed network speed in the beginning. Replacing the switching fabric will put us where we should be on the list."

The fastest high-performance clustering interconnects are devices such as Myrinet, made by Myricom Inc. However, these interconnects are usually expensive—about $1,000 a pop (or $300,000 for a typical HPCC) and $100,000 more for the Myrinet switch. This is too pricey for most academic concerns, but if Stanford had gone with

Gfunk

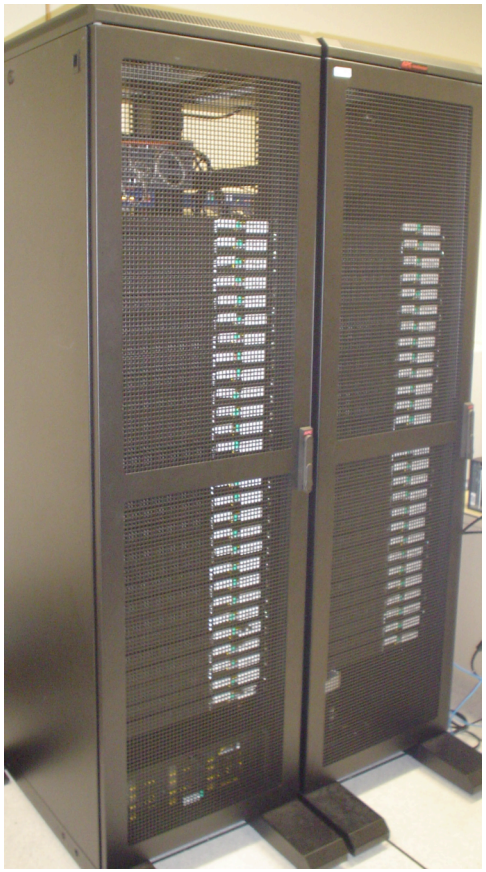
- 164 Processor Intel Xeon 2.8GHz
- Myrinet
- Fast Ethernet
- 1TB NFS Appliance

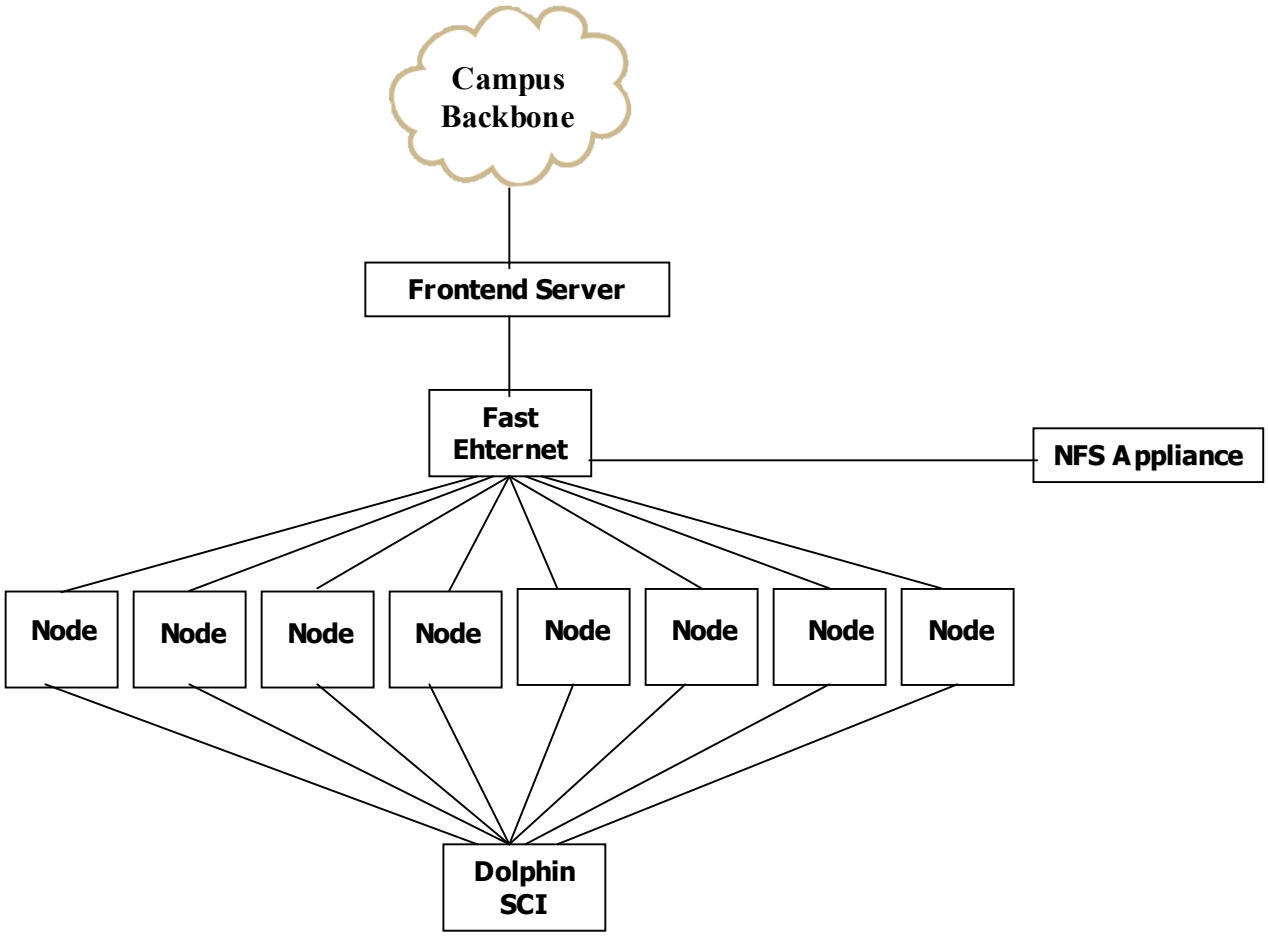

Campus
Backbone
Frontend Server
GigE Net
NFS Appliance
Node
Node
Node
Node
Node
Node
Node
Node
Myrinet

Firn

- 112 Processor Intel Pentium 1.0GHz
- 1GB RAM per node
- Fast Ethernet
- Dolphin SCI

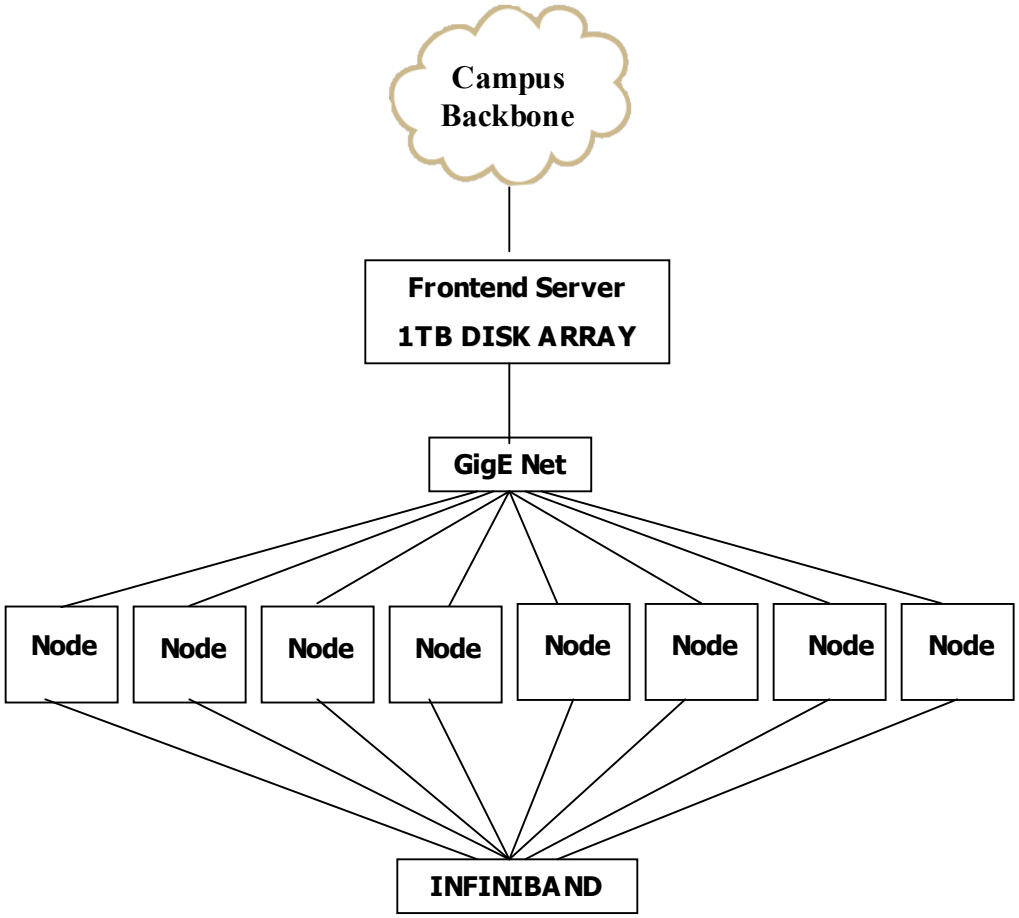
  (in progress.. in my spare time)




Campus
Backbone
Frontend Server
Fast
Ehternet
NFS Appliance
Node
Node
Node
Node
Node
Node
Node
Node
Dolphin
SCI

Sintering

- 48 processor AMD Opteron
- Infiniband
- Gigabit Ethernet
- 1TB on Frontend


Campus
Backbone
Frontend Server
1TB DISK ARRAY
GigE Net
Node
Node
Node
Node
Node
Node
Node
Node
INFINIBAND

Nivation

- 164 Processor Intel Xeon 3.0GHz
- 4GB RAM per node
- Myrinet
- Gigabit Ethernet
- 2-1TB NAS Appliance
- 4 Tools Nodes

Eliminated Bottlenecks

Redundancy

400MBytes/sec

Huge Bottleneck/

Single Point of Failure

Panasas Integration

•Installation and configuration of Panasas Shelf

•Switch configuration changes for link aggregation

•Copy RPM to /home/install/contrib/enterprise/3/public/i386/RPMS

•create/edit extend-compute.xml:

```
# Add panfs to fstab
REALM=10.10.10.10
mount_flags="rw,noauto,panauto"
/bin/rm -f /etc/fstab.bak.panfs
/bin/rm -f /etc/fstab.panfs
/bin/cp /etc/fstab /etc/fstab.bak.panfs
/bin/grep -v "panfs://" /etc/fstab > /etc/fstab.panfs
/bin/echo "panfs://$REALM:global  /panfs  panfs  $mount_flags      0      0" >> /etc/fstab.panfs
/bin/mv -f /etc/fstab.panfs /etc/fstab
/bin/sync

/sbin/chkconfig --add panfs
/usr/local/sbin/check_panfs
LOCATECRON=/etc/cron.daily/slocate.cron
LOCATE=/etc/sysconfig/locate
LOCTEMP=/tmp/slocate.new

/bin/cat $LOCATECRON | sed "s/,proc,/,proc,panfs,/g" > $LOCTEMP
/bin/mv -f $LOCTEMP $LOCATECRON
/bin/cat $LOCATECRON | sed "s/\/afs,/\/afs,\/panfs,/g" > $LOCTEMP
/bin/mv -f $LOCTEMP $LOCATECRON
```

•[root@rockscluster]# rocks-dist dist ; cluster-fork '/boot/kickstart/cluster-kickstart'

•/etc/auto.home userX -fstype=panfs panfs://10.x.x.x/home/userX

Benchmarking Panasas

```
#!/bin/bash

#PBS -N BONNIE

#PBS -e Log.d/BONNIE.panfs.err

#PBS -o Log.d/BONNIE.panfs.out

#PBS -m aeb

#PBS -M stevejones@stanford.edu

#PBS -l nodes=1:ppn=2

#PBS -l walltime=30:00:00

PBS_O_WORKDIR='/home/smjones/benchmarks'
export PBS_O_WORKDIR

### ---------------------------------------
### BEGINNING OF EXECUTION
### ---------------------------------------

echo The master node of this job is `hostname`
echo The job started at `date`
echo The working directory is `echo $PBS_O_WORKDIR`
echo This job runs on the following nodes:
echo `cat $PBS_NODEFILE`

### end of information preample

cd $PBS_O_WORKDIR
cmd="/home/tools/bonnie++/sbin/bonnie++ -s 8000 -n 0 -f -d /home/smjones/bonnie"
echo "running bonnie++ with: $cmd in directory "`pwd`
$cmd >& $PBS_O_WORKDIR/Log.d/run9/log.bonnie.panfs.$PBS_JOBID
```

NFS - 8 Nodes

Version 1.03	------Sequential Output------				--Sequential Input-		--Random-			
	-Per Chr-	--Block--	-Rewrite-		-Per Chr-	--Block--	--Seeks--			
Machine	Size K/sec	%CP	K/sec	%CP	K/sec	%CP	K/sec	%CP	K/sec	%CP
compute-1-26.	8000M	6146	4	3554	2	17072	3	350.2	1	/sec %CP
compute-1-25.	8000M	6023	4	3545	2	17145	3	287.4	1	
compute-1-24.	----------									
compute-1-23.	----------									
compute-1-22.	----------									
compute-1-21.	----------									
compute-1-20.	----------									
compute-1-19.	----------									

PanFS - 8 Nodes

Version 1.03	------Sequential Output------						--Sequential Input-		--Random-	
	-Per Chr-	--Block--	-Rewrite-		-Per Chr-	--Block--			--Seeks--	
Machine	Size	K/sec %CP	K/sec	%CP	K/sec	%CP	K/sec	%CP	K/sec	%CP
compute-1-18.	8000M	20767 8	4154	3			24460	7	72.8	0
compute-1-17.	8000M	19755 7	4009	3			24588	7	116.5	0
compute-1-16.	8000M	19774 7	4100	3			23597	7	96.4	0
compute-1-15.	8000M	19716 7	3878	3			25384	8	213.6	1
compute-1-14.	8000M	19674 7	4216	3			24495	7	72.8	0
compute-1-13.	8000M	19496 7	4236	3			24238	7	71.0	0
compute-1-12.	8000M	19579 7	4117	3			23731	7	97.1	0
compute-1-11.	8000M	19688 7	4038	3			24195	8	117.7	0

154MB/sec for concurrent write

NFS - 16 Nodes

Version 1.03	------Sequential Output------				--Sequential Input-		--Random-			
	-Per Chr-	--Block--	-Rewrite-		-Per Chr-		--Block--		--Seeks--	
Machine	Size	K/sec %CP	K/sec	%CP	K/sec	%CP	K/sec	%CP	/sec	%CP
compute-1-19.	8000M	8354 6	3487	2			14072	3	296.5	1
compute-1-18.	8000M	4505 2	4174	3			30380	6	862.8	3
compute-1-17.	----------									
compute-1-16.	----------									
compute-1-15.	----------									
compute-1-14.	----------									
compute-1-13.	----------									
compute-1-12.	----------									
compute-1-11.	----------									
compute-1-10.	----------									
compute-1-9.	----------									
compute-1-8.	----------									
compute-1-7.	----------									
compute-1-6.	----------									
compute-1-5.	----------									
compute-1-4.	----------									

PanFS - 16 Nodes

Version 1.03		------Sequential Output------				--Sequential Input-		--Random-	
		-Per Chr-		--Block--		-Rewrite-		-Per Chr-		--Block--		--Seeks--	
Machine	Size	K/sec	%CP	K/sec	%CP	K/sec	%CP	K/sec	%CP	K/sec	%CP	/sec	%CP
compute-1-26.	8000M			14330	5	3392	2			28129	9	54.1	0
compute-1-25.	8000M			14603	5	3294	2			30990	9	60.3	0
compute-1-24.	8000M			14414	5	3367	2			28834	9	55.1	0
compute-1-23.	8000M			9488	3	2864	2			17373	5	121.4	0
compute-1-22.	8000M			8991	3	2814	2			21843	7	116.5	0
compute-1-21.	8000M			9152	3	2881	2			20882	6	80.6	0
compute-1-20.	8000M			9199	3	2865	2			20783	6	85.2	0
compute-1-19.	8000M			14593	5	3330	2			29275	9	61.0	0
compute-1-18.	8000M			9973	3	2797	2			18153	5	121.6	0
compute-1-17.	8000M			9439	3	2879	2			22270	7	64.9	0
compute-1-16.	8000M			9307	3	2834	2			21150	6	99.1	0
compute-1-15.	8000M			9774	3	2835	2			20726	6	77.1	0
compute-1-14.	8000M			15097	5	3259	2			32705	10	60.6	0
compute-1-13.	8000M			14453	5	2907	2			36321	11	126.0	0
compute-1-12.	8000M			14512	5	3301	2			32841	10	60.4	0
compute-1-11.	8000M			14558	5	3256	2			33096	10	62.2	0

187MB/sec for concurrent write

Capacity imbalances on jobs - only 33MB/sec increase from 8 to 16 job runs…

PanFS - 16 Nodes

[pancli] sysstat storage

IP	CPU Util	Disk Util	Ops/s In	KB/s Out	Capacity (GB) Total	Avail	Reserved	
10.10.10.250	55%	22%	127	22847	272	485	367	48
10.10.10.253	60%	24%	140	25672	324	485	365	48
10.10.10.245	53%	21%	126	22319	261	485	365	48
10.10.10.246	55%	22%	124	22303	239	485	366	48
10.10.10.248	57%	22%	134	24175	250	485	369	48
10.10.10.247	52%	21%	124	22711	233	485	366	48
10.10.10.249	57%	23%	135	24092	297	485	367	48
10.10.10.251	52%	21%	119	21435	214	485	366	48
10.10.10.254	53%	21%	119	21904	231	485	367	48
10.10.10.252	58%	24%	137	24753	300	485	366	48
Total "Set 1"	55%	22%	1285	232211	2621	4850	3664	480

Sustained BW 226 MBytes/Sec during 16 1GB concurrent writes

PanFS - 16 Nodes

[pancli] sysstat storage

IP	CPU Util	Disk Util	Ops/s In	KB/s Out	Capacity (GB) Total	Avail	Reserved	
10.10.10.250	58%	95%	279	734	21325	485	355	48
10.10.10.253	60%	95%	290	727	22417	485	353	48
10.10.10.245	54%	92%	269	779	19281	485	353	48
10.10.10.246	59%	95%	290	779	21686	485	354	48
10.10.10.248	60%	95%	287	729	22301	485	357	48
10.10.10.247	52%	91%	256	695	19241	485	356	48
10.10.10.249	57%	93%	276	708	21177	485	356	48
10.10.10.251	49%	83%	238	650	18043	485	355	48
10.10.10.254	45%	82%	230	815	15225	485	355	48
10.10.10.252	57%	94%	268	604	21535	485	354	48
Total "Set 1"	55%	91%	2683	7220	202231	4850	3548	480

Sustained BW 197 MBytes/Sec during 16 1GB concurrent sequential reads


```
[pancli] sysstat storage
IP          CPU   Disk  Ops/s  KB/s       Capacity (GB)
            Util  Util    In    Out  Total  Avail  Reserved
10.10.10.250   6%   5%    35   292   409    485    370      48
10.10.10.253   5%   4%    35   376   528    485    368      48
10.10.10.245   4%   3%    29   250   343    485    368      48
10.10.10.246   6%   4%    28   262   373    485    369      48
10.10.10.248   5%   3%    27   234   290    485    372      48
10.10.10.247   3%   3%     1     1     2    485    370      48
10.10.10.249   5%   3%    48   258   365    485    371      48
10.10.10.251   4%   3%    46   216   267    485    369      48
10.10.10.254   4%   3%    32   256   349    485    370      48
10.10.10.252   4%   3%    34   337   499    485    370      48
Total "Set 1"  4%   3%   315  2482  3425   4850   3697     480
```

Average sustained BW 2.42 Mbytes/sec in - 3.34 Mbytes/sec out at 76% +/- CPU utilization

```
[root@frontend-0 root]# showq
ACTIVE JOBS--------------------
JOBNAME          USERNAME      STATE  PROC   REMAINING            STARTTIME

8649              smjones    Running     2  1:04:55:08  Sun May 15 18:20:23
8660              boumosle   Running     6  1:23:33:15  Sun May 15 18:58:30
8524               arallu    Running    16  2:01:09:51  Fri May 13 20:35:06
8527               arallu    Running    16  2:01:23:19  Fri May 13 20:48:34
8590                rajay    Running    64  3:16:42:50  Sun May 15 10:08:05
8656                rajay    Running    16  4:00:55:36  Sun May 15 18:20:51
8647              cdastill   Running     5 99:22:50:42  Sun May 15 18:15:58

   7 Active Jobs     125 of  164 Processors Active (76.22%)
                      65 of   82 Nodes Active      (79.27%)
```

Center for Integrated Turbulence Simulations

Managing and Maintaining Systems

Managing Hardware

- Tool-less parts replacement
- Location of parts fulfillment center
- Service Contract
  - 4 hour response on critical components
  - Next business day on compute nodes

Managing Hardware

- Maintain cabling standards
- Accessibility
- Ease of identification
- Ease of replacement

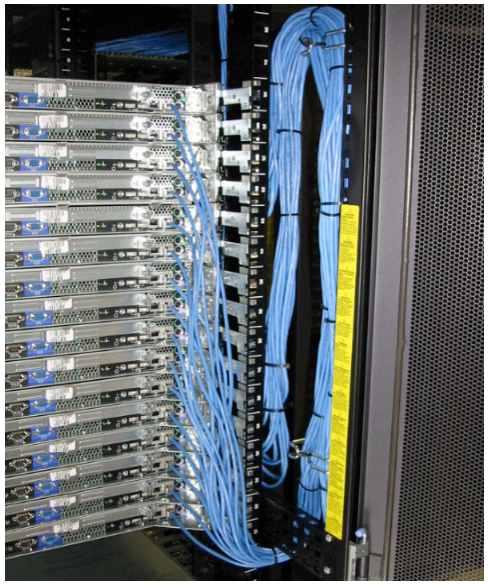
Our New Cluster Ordering Methodology

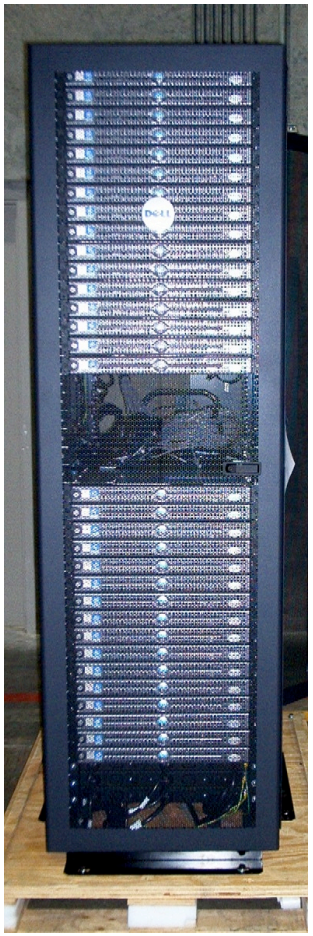
Merge Center

Let the vendor:

- Unpack the boxes
- Label the nodes and cables
- Pack it up
- Ship it for installation

Merge Center





Merge Center

Without Rack & Stack

- **Product arrives at various times**
- **Services performed on-site by customer or 3rd party**
- **4-6 week cycle time from order entry to installation**
- **Variable quality and standards**

With Rack & Stack

- **Direct ship to the customer site**
- **Product arrives in a single integrated shipment**
- **2-3 week cycle time from order entry**
- **Standard quality integration, cabling and labeling**

-Support Rocks -

1. Register Your Cluster

- (65) Iceberg Bio-X @ Stanford University Pentium 4 604 2.80 3382.4 Stanford, CA
- (367) Firn ICME @ Stanford University Pentium 3 112 1.00 112 Stanford, CA
- (368) Sintering ICME @ Stanford University Opteron 48 1.60 153.6 Stanford, CA
- (369) Regelation ICME @ Stanford University Pentium 4 8 3.06 48.96 Stanford, CA
- (370) GFunk ICME @ Stanford University Pentium 4 164 2.66 872.48 Stanford, CA
- (301) Nivation ICME @ Stanford University Pentium 4 172 3.06 1052.64 Palo Alto, CA

2. Demand your vendors support it

Questions?

Steve Jones stevejones@stanford.edu

http://www.hpcclusters.org