Some things work…
Some things could be better…
What lies ahead…

Stephen L. Scott

scottsl@ornl.gov

www.csm.ornl.gov/~sscott

---

presented at:

SOS 7

March 5, 2003

Oak Ridge National Laboratory

my motivation for this discussion…

- Thursday – Feb 27 (Ron Brightwell)
  - I think it would be good to have your perspective on what works, what doesn't, and what is needed is for current and future large-scale supercomputers. I'll let you pick the area that you'd like to focus on. People generally have an area that they feel strongly about, so I'd like to hear yours.

- Reminder to me…
  - all clusters are not supercomputers
    - proof is exercise left to the viewer…
  - all supercomputers will be (are) clusters
    - no matter how large a monolithic machine is, we will network a bunch together
  - never use absolutes…

Some things work...

Open Source Cluster Application Resources

What does OSCAR do?

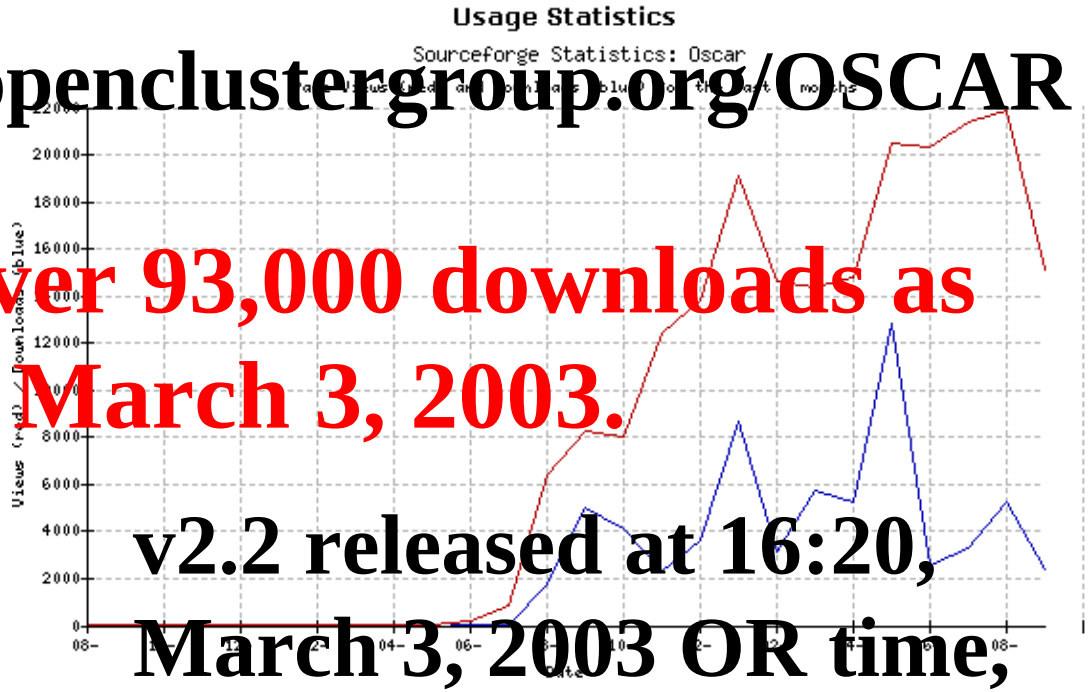
- Wizard based cluster software installation
  - Operating system
  - Cluster environment
- Automatically configures cluster components
- Increases consistency among cluster builds
- Reduces time to build / install a cluster
- Reduces need for expertise

www.openclustergroup.org/OSCAR

OSCAR, over 65,000 customers served!

- oscar.sourceforge.net
  - 65,936 downloads
  - 121,355 page hits

(September 22, 2002 – 12:30am)



Over 93,000 downloads as of March 3, 2003.

v2.2 released at 16:20, March 3, 2003 OR time, in 7hr 40min

991 page hits

120 downloads

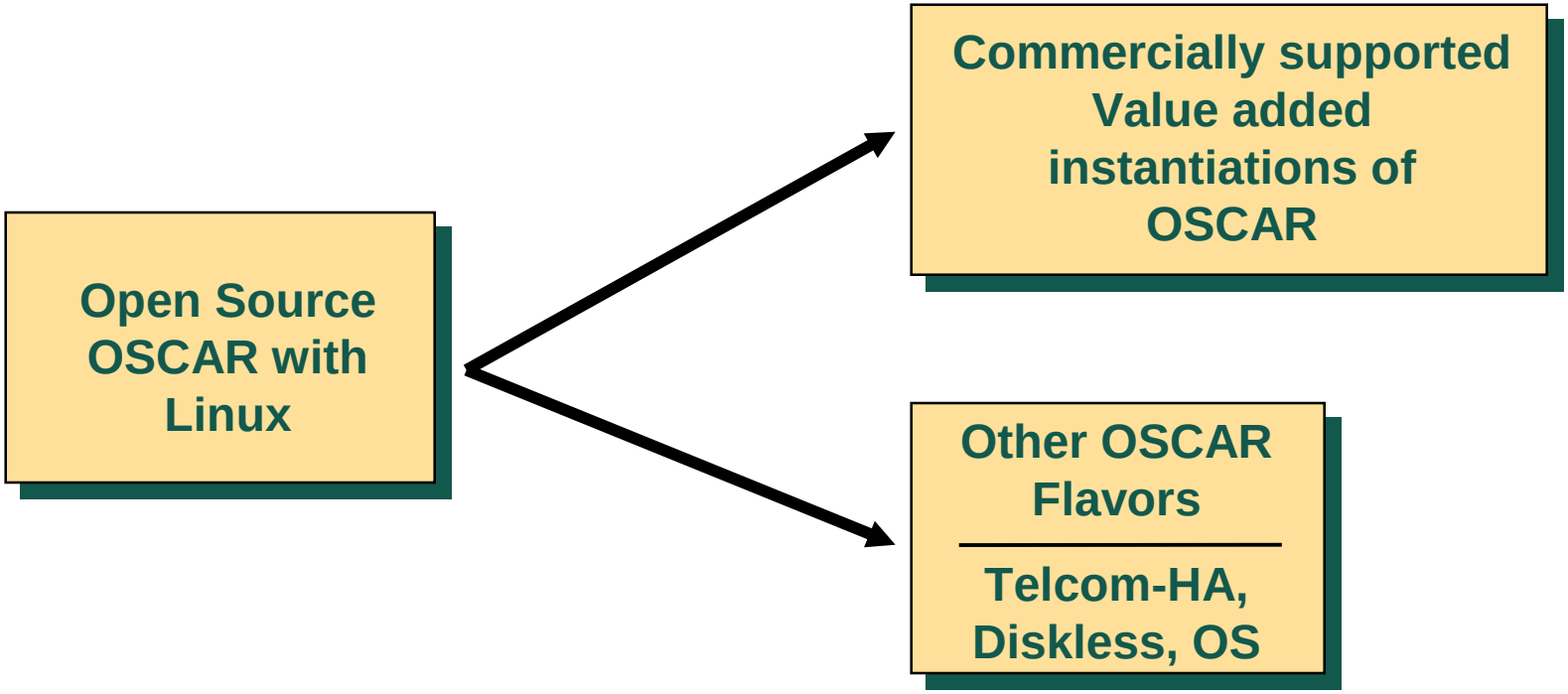
Month	Rank	Page Views	D/l	Bugs	Support	Patches	All Trkr	Tasks	CVS
September 2002	1 (0.00)	6,138	2,[illegible]	14 (21)	0 (0)	0 (0)	20 (23)	0 (0)	104
August 2002	78 ([illegible])	[illegible]	[illegible]	56 (56)	0 (0)	0 (0)	65 (62)	0 (2)	258
July 2002	11 ([illegible])	[illegible]	[illegible]	46 (46)	0 (0)	0 (0)	54 (59)	2 (6)	239
June 2002	128 (98.58)	20,302	2,561	39 (31)	0 (0)	0 (0)	45 (31)	0 (1)	165
May 2002	217 (97.60)	20,499	12,843	38 (25)	0 (0)	0 (0)	39 (38)	0 (0)	142
April 2002	2[illegible]	[illegible]	[illegible]	[illegible]	0 (1)	0 (0)	20 (27)	0 (0)	137
March 2002	41[illegible]	[illegible]	[illegible]	[illegible]	0 (0)	0 (0)	19 (7)	0 (0)	18
February 2002	320 (64.62)	14,[illegible]	3,122	16 (12)	2 (2)	0 (0)	28 (15)	0 (0)	66
January 2002	306 (94.81)	19,069	8,696	32 (31)	0 (2)	0 (0)	34 (33)	0 (0)	100
December 2001	298 ([illegible])	13,797	3,598	36 (23)	[illegible]	0 (0)	40 (23)	0 (0)	156
November 2001	102 ([illegible])	[illegible]	[illegible]	[illegible]	[illegible]	0 (0)	2 (1)	0 (0)	2
October 2001	59[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0 (0)	0 (0)	0 (0)	61
September 2001	587 (90.02)	8,244	5,006	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	10
August 2001	574 (90.84)	6,396	1,744	3 (10)	0 (0)	0 (0)	3 (10)	0 (0)	89
July 2001	2233 (62.39)	829	19	11 (8)	0 (0)	0 (0)	13 (8)	0 (0)	106
June 2001	0 (0.00)	211	7	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	3
May 2001	0 (0.00)	28	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0
April 2001	0 (0.00)	4	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0
March 2001	0 (0.00)	4	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0
February 2001	0 (0.00)	9	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0
January 2001	0 (0.00)	13	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0
December 2000	0 (0.00)	19	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0
November 2000	0 (0.00)	1	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0
October 2000	0 (0.00)	1	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0
September 2000	0 (0.00)	3	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0
August 2000	365 (55.78)	3	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0

Statistics for All Time

Lifespan	Rank	Page Views	D/l	Bugs	Support	Patches	All Trkr	Tasks	CVS
days	474 (88.57)	212,137	65,936	319 (303)	5 (5)	0 (0)	382 (337)	2 (9)	1,656

The OSCAR strategy

- OSCAR is a snap-shot of best-known-methods for building, programming and using clusters of a “reasonable” size.
- To bring uniformity to clusters, foster commercial versions of OSCAR, and make clusters more broadly acceptable.
- Consortium of research, academic & industry members cooperating in the spirit of open source.



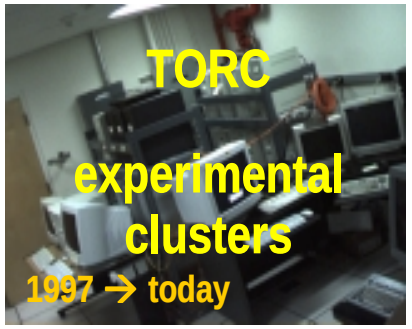
Functional Areas

- **cluster installation** – System Installer Suite (SIS)
- **programming environment** – PVM, LAM/MPI, MPICH
- **workload management** – MAUI, PBS
- **security** – SSH, restricted compute nodes
- **administration** – Cluster Command & Control (C3)
- **maintenance** – C3, SIS, Switcher, Ganglia
- **documentation**
- **packaging** – including: post install test of cluster as a system

OSCAR is effectively a Beowulf linux cluster distro

ORNL Research Clusters

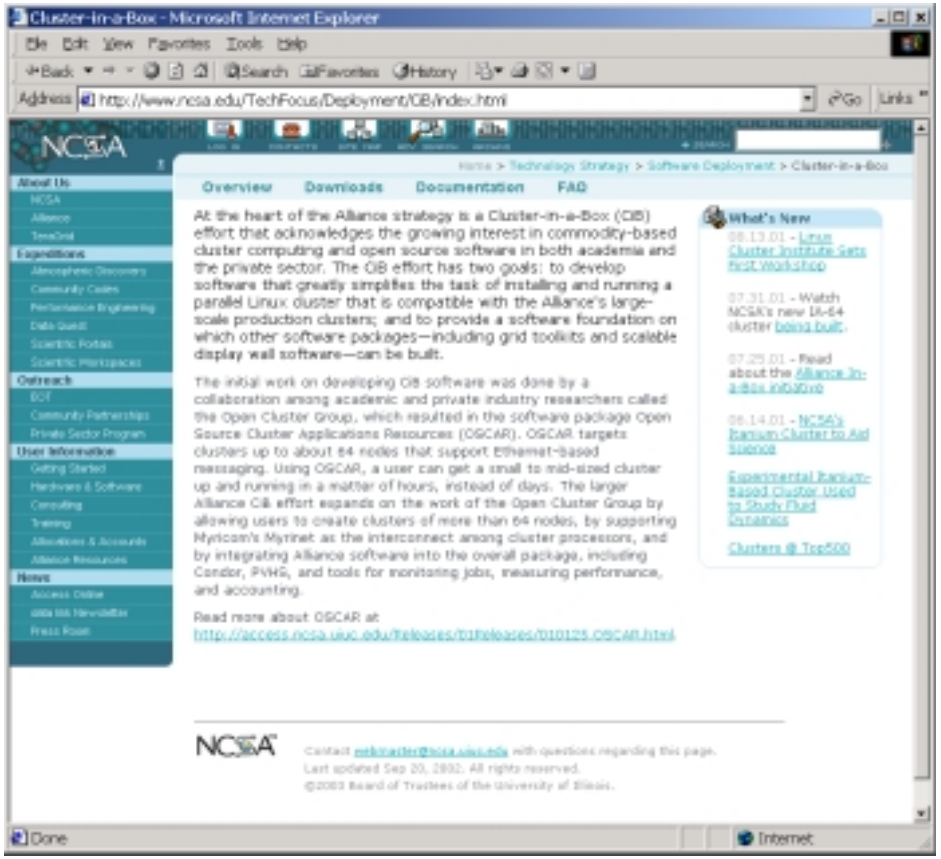
- TORC – Tennessee Oak Ridge Cluster
  - Component based performance evaluation
  - Cluster test bed
  - Heterogeneous Cluster test bed
- HighTORC & eXtremeTORC
  - Scalability test bed for SciDAC SSS
  - COTS Cluster HPC test bed
  - Gigabit cluster network test bed
  - Applications
    - Visualization over the Grid
    - Climate modeling
    - Genomics
    - High energy physics
    - Statistical modeling

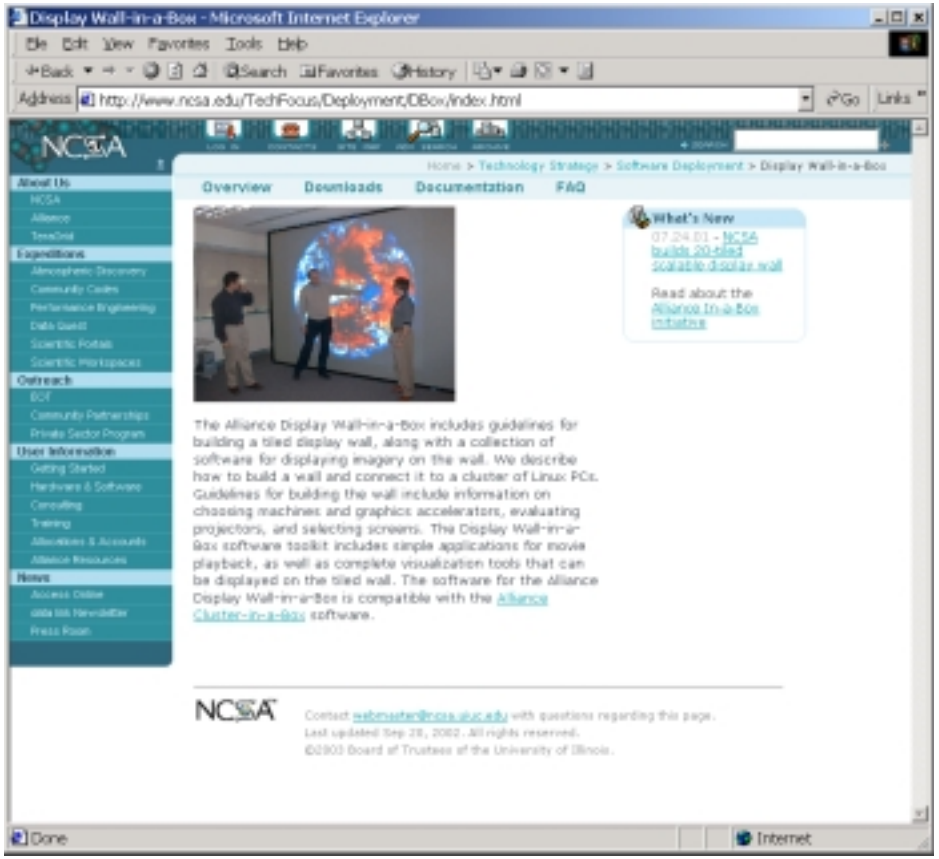


Commercial clusters

- Companies selling / loading clusters with OSCAR (that we know of…)
  - MSC.Linux
  - RackSaver
  - EbizEnterprises
  - IBM
  - Dell
  - LinuxNetworx

NCSA: Cluster in-a-Box -and- Display wall-in-a-Box





Some things could be better…

Building cluster distro multiplies impact of component faults

- support for multiple versions of each distro
  - some things fixed are subsequently broken
  - some things broken are subsequently fixed
    - RPM broken in rh7.2 / fixed in rh7.3
    - rh7.0
  - libraries change and some packages must be rebuilt
    - rh8.0 will have new 3.x GCC
      - libs/rpms recompile/rebuild for shared lib compatibility

Building cluster distro multiplies impact of component faults

- support for multiple distros
  - distro-centric commands, parameters, switches, etc…
    - configuration commands
      - turn off sendmail
        - » root# service sendmail stop
      - disable sendmail to prevent autostartup on boot
        - » root# chkconfig --level 2345 sendmail off
  - different naming conventions
    - MySQL
      - RH7.3 mysql-3.23.49-3.i386.rpm
      - MDK8.2 MySQL-3.23.47-5mdk.i586.rpm

Building cluster distro multiplies impact of component faults

- Whose fault are these…
  - Python – (systems live on python code…)
    - rh7.x systems included either
      - Python: v1.5 (rh7.1, 7.2, 7.3)
      - Python2: v1.5 and v2.x (rh7.2, 7.3)
    - caused confusion with packages
    - workaround
      - supply own Python2 package for rh7.1
      - make sure not to clobber Python2 that already exists in rh7.2 and 7.3

Did anyone think to ask if it works?

- how to test an RPM
  - did it install?

- comprehensive cluster distro testing

  across versions X across distros = limited distro/version support

What lies ahead…

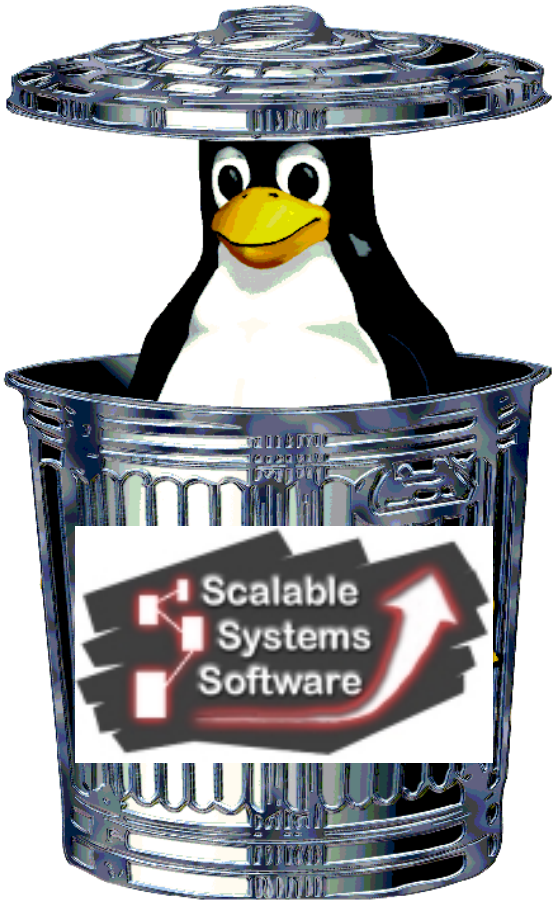
Will someone save us?

- We need some distro / configuration conventions

- Linux Standard Base (LSB)
- United Linux (shrinking marketplace?)
  - Conectiva
  - SCO
  - SuSE
  - Turbo
- Special HPC kernel build for compute nodes
  - much interest – but…
    - who will fund this?
    - who will support this?
      - can we support this effort in long term?

What you will see out of Oak Ridge

- configuration management
  - cluster build
  - node management
  - package management
  - administration
- resource management
  - scheduling
  - queue manager
  - accounting
- process management
  - process launching
  - state management
  - checkpoint / restart



ORNL
ANL
LBNL
AMES
PNNL
SNL
LANL
NCSA

SciDAC
Scientific Discovery through Advanced Computing

DOE Office of Science ASCR BER BES FE HENP MICS

Questions?