

The OSCAR Solution Stack for Cluster Computing

The OSCAR Solution Stack for Cluster Computing

Tim Mattson

Intel Corp
Computational Software labs
U.S.A.

timothy.g.mattson@intel.com



OSCAR

 Skip my intro

Agenda

- ➡ • Cluster Computing Software Stacks
 - What they are and why we desperately need a common base software stack.
- OSCAR and the Open Cluster Group
- OSCAR Today
- OSCAR Tomorrow
- Cluster Computing Standards

The OSCAR Solution Stack for Cluster Computing

Intel wants to get rich with clusters

The role of Independent Software Vendors

- Intel likes cluster since they so many high-end processors per system.
- ... but people don't buy computers, they buy problem solutions (i.e. applications+computers to run them).
- Hence if we want a healthy HPC industry, we need a rich set of cluster-enabled applications.

The academic/national lab shares this need with industry.
Industrial HPC use translates into better products and more funding

So where is the cluster software?

- Except for a few CFD[#] codes, crash codes, a handful of bioinformatics and chemistry codes, and a smattering of other codes, there aren't many ISV* supported cluster enabled codes.
- Most ISV's have ignored parallel computing.
- Why is this the case?

*ISV = Independent Software Vendor

[#]CFD = Computational Fluid Dynamics

The OSCAR Solution Stack for Cluster Computing

In the late 80's, we thought portable programming environments were the solution.

So computer scientists created a few...

ABCPL	CORRELATE	GLU	Mentat	Parafuse2	pC++
ACE	CPS	GUARD	Legion	Paralation	SCHEDULE
ACT++	CRL	HAsL	Meta Chaos	Parallel-C++	ScTL
Active messages	CSP	Haskell	Midway	Parallaxis	POET
Adl	Cthreads	HPC++	Millipede	ParC	SDDA
Adsmith	CUMULVS	JAVAR	CparPar	ParLib++	SHMEM
ADDAP	DAGGER	HORUS	Mirage	ParLin	SIMPLE
AFAPI	DAPPLE	HPC	MpC	Parmacs	Sina
ALWAN	Data Parallel C	IMPACT	MOSIX	Parti	SISAL
AM	DC++	ISIS	Modula-P	pC	distributed smalltalk
AMDC	DCE++	JAVAR	Modula-2*	pC++	SMI
AppLeS	DDD	JADE	Multipol	PCN	SONIC
Amoeba	DICE	Java RMI	MPI	PCP:	Split-C
ARTS	DIPC	javaPG	MPC++	PH	SR
Adhupascan-db	DOLIB	JavaSpace	Murin	PEACE	Sthreads
Aurora	DOME	JIDL	Nano-Threads	PCU	Strand
Autonap	DOSMOS	Joyce	NESL	PET	SUIF
bb_threads	DRL	Khoros	NetClasses++	PETSc	Synergy
Blaze	DSM-Threads	Karma	Nexus	PENNY	Telegraphos
BSP	Ease	KOAN/Fortran-S	Nimrod	Phosphorus	SuperPascal
BlockComm	ECO	LAM	NOW	POET	TCGMSG
C*	Eiffel	Lilac	Objective Linda	Polaris	Threads.h++
C* in C	Eilean	Linda	Occam	POOMA	TRAPPER
C**	Emerald	JADA	Omega	POOL-T	uC++
CarlOS	EPL	WWWinda	OpenMP	PRESTO	UNITY
Cashmere	Excilbur	SETL-Linda	Orca	P-RIO	UC
C4	Express	ParLin	OOF90	Prospero	V
CC++	Falcon	Eilean	P++	Proteus	VIC*
Chu	Filaments	P4-Linda	P3L	QPC++	Visifold V-NUS
Charlotte	FM	Glenda	p4-Linda	PVM	VPE
Cham	FLASH	POSYBL	Pablo	PSI	Win32 threads
Charm++	The FORCE	Objective-Linda	PADE	PSDM	WWWinda
Cid	Fork	LIPS	PADRE	Quake	XENOOOPS
Clk	Fortran-M	Locust	Panda	Quark	XPC
CM-Fortran	FX	Lpars	Papers	Quick Threads	Zounds
Converse	GA	Lucid	GA	Sage++	ZPL
Code	GAMMA	Maisie	Para++	SCANDAL	
COOL	Glenda	Manifold	Paradigm	SAM	

And the
ISV's
laughed
at us.

“Too
many”
options
don't
help.

In the mid-90's, standard API's were the solution.

So we finally agreed on a small set of API's...

- Thread Libraries
 - Win32 API
 - POSIX threads.
- Compiler Directives
 - OpenMP - portable shared memory parallelism.
- Message Passing Libraries
 - MPI - message passing

We picked up a few more ISV's, but most ignored us.

The OSCAR Solution Stack for Cluster Computing

What's missing?

**A standard API isn't enough ... ISV's
need a standard platform to build
their business upon.**

**... and the platform must be easy for
the general user to use.**

Cluster Computing Platforms

- **A Platform is the API's, tools, interfaces, and everything else required to install, maintain and use a computing system.**

It includes:

- | | |
|-------------------------|--------------------------|
| • Sys admin tools | • Batch queue |
| • Installers | • Scheduler |
| • "Parallel unix tools" | • Performance monitoring |
| • Libraries | • Debugging |

The OSCAR Solution Stack for Cluster Computing

So we all ran off and created cluster platforms

Beowulf “how to” Linux Networkx
Extreme Linux

 Alinka NPACI-Rocks
Score SE

... and countless proprietary and home-brewed platforms.

Are we at risk of scaring away the ISV's
(just as we did with API-glut in the early 90's)?

*Brands and names are property of their respective owners.

My hope is ...

- To see a small number (on the order of two) common base platforms emerge.
 - ◆ Look very similar to end-users.
 - ◆ “Vendors” distinguish themselves with “value added instantiations”, not new interfaces.
- To try and make this happen, a group of us came together as “The Open Cluster Group”.

The OSCAR Solution Stack for Cluster Computing

Agenda

- Cluster Computing Software Stacks
 - What they are and why we desperately need a common base software stack.
- • OSCAR and the Open Cluster Group
 - OSCAR Today
 - OSCAR Tomorrow
 - Cluster Computing Standards

Cluster Software stacks

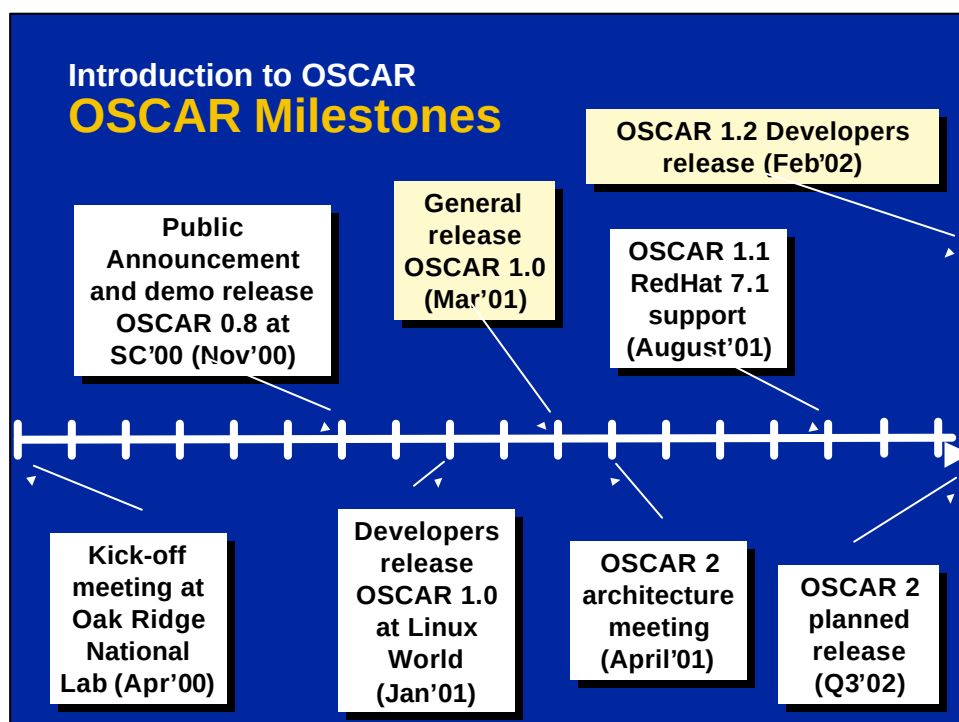
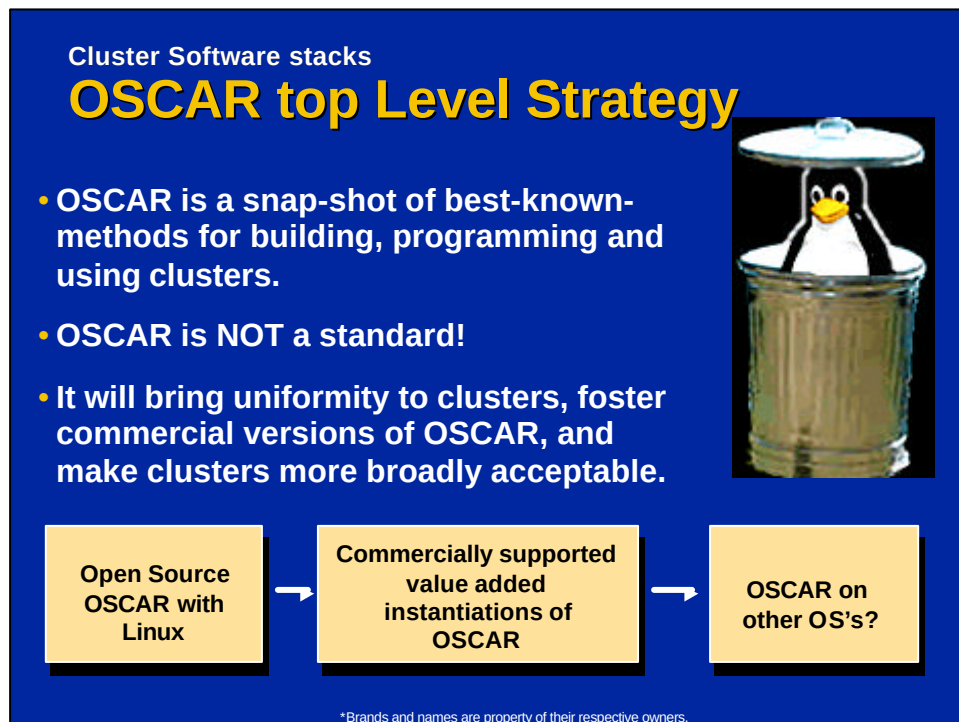
The Open Cluster Group

- We are an open group dedicated to open source solutions for cluster computing.
 - ◆ Dell, IBM, Intel, LLNL, MSC.software, NCSA, ORNL, SGI, Indiana University and others.
- Our first project is OSCAR
 - ◆ Open Source Cluster Application Resources

Follow our progress at:
<http://www.openclustergroup.org>

*Brands and names are property of their respective owners.

The OSCAR Solution Stack for Cluster Computing



The OSCAR Solution Stack for Cluster Computing

OSCAR:

How are we doing?

- OSCAR is doing great:
 - “We’re number 1”.
 - Over 16,000 downloads in 2001.
 - 35% in a recent poll on www.cluster.top500.org
 - Lively on-line user group – starting to answer questions before we can get to them!
 - OSCAR has brought many new people into clustering
- OSCAR is not doing very well:
 - Maintenance mode still not enabled.
 - Poor support for adding/deleting nodes.
 - Too hard to adjust to new releases from RedHat.
 - OSCAR is still way to complicated.

None of us are satisfied. OSCAR isn't ready for the production oriented “commercial world”

(I send such people to Scyld or MSC.software).

Agenda

- Cluster Computing Software Stacks
 - What they are and why we desperately need a common base software stack.
- OSCAR and the Open Cluster Group
- ➡ • OSCAR Today
- OSCAR Tomorrow

The OSCAR Solution Stack for Cluster Computing

OSCAR Basics

- **Version 1.0, 1.1**
 - ◆ LUI = Linux Utility for cluster Install
 - Network boots nodes via PXE or floppy
 - Nodes install themselves from rpms over NFS from the server
 - Post installation configuration of nodes and server executes
- **Version 1.2+**
 - ◆ SIS = System Installation Suite
 - System Imager + LUI = SIS
 - Creates image of node filesystem locally on server
 - Network boots nodes via PXE or floppy
 - Nodes synchronize themselves with server via rsync
 - Post installation configuration of nodes and server executes

Introduction to OSCAR

OSCAR Contents/Status

OS Layer	OSCAR 1.0: RedHat 6.2 OSCAR 1.2: RedHat 7.2
Installation & System Configuration	OSCAR 1.0, 1.1: LUI OSCAR 1.2 and on: SIS
Security	OpenSSH / OpenSSL
System Services & Cluster Management	C3
Job Management	PBS, Maui scheduler.
Programming Environment	gcc, PVM, MPIch and LAM-MPI
Packaging	Components integrated and auto-installed. Documentation.

*Brands and names are property of their respective owners.

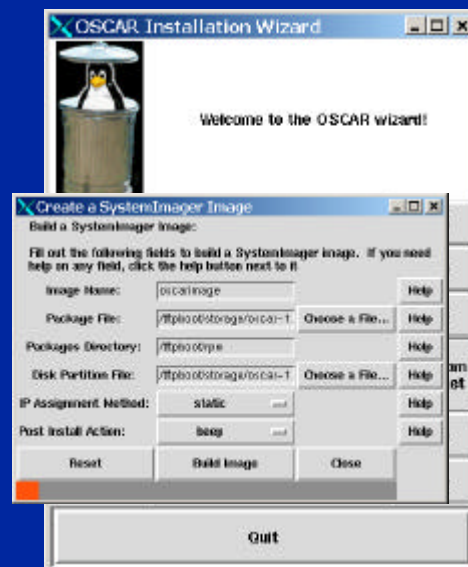
The OSCAR Solution Stack for Cluster Computing

Hardware Considerations

- **Server & Clients**
 - ◆ Must be IA32 systems – IA64 in pre-beta.
 - ◆ Must be connected by an Ethernet network (preferably a private one)
- **Clients**
 - ◆ Should contain identical hardware
 - ◆ PXE Enabled NIC or Floppy Drive

Installation Overview

- Install RedHat 7.2
- Download OSCAR
- Copy RPMS to server
- Run wizard
 - ◆ Build image per client type (partition layout, HD type)
 - ◆ Define clients (network info, image binding)
 - ◆ Setup networking (collect MAC addresses, configure DHCP, build boot floppy)
 - ◆ Boot clients / build
 - ◆ Complete setup (post install)
 - ◆ Install test suite



The OSCAR Solution Stack for Cluster Computing

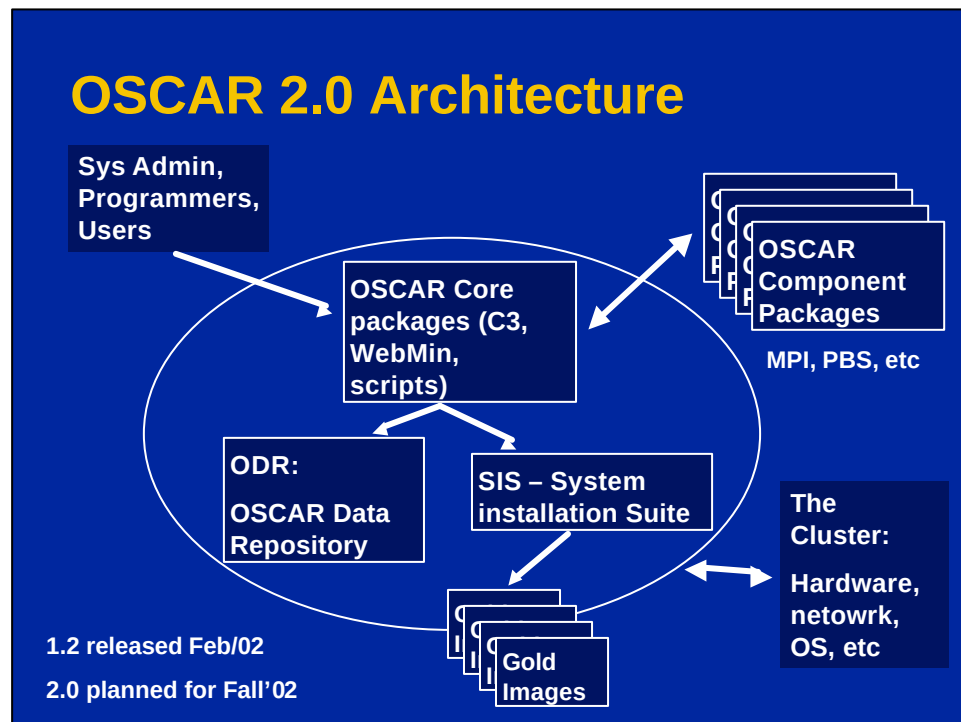
Agenda

- Cluster Computing Software Stacks
 - What they are and why we desperately need a common base software stack.
- OSCAR and the Open Cluster Group
- OSCAR Today
- ➡ ● OSCAR Tomorrow
- Cluster Computing Standards

What OSCAR needs

- Movement away from a reliance on RedHat.
- Greatly improved ease of use – a better wizard, more automation.
- Support for huge clusters.
- Tools to support cluster maintenance:
 - ◆ Adding and deleting nodes.
 - ◆ Adding and deleting software packages.

The OSCAR Solution Stack for Cluster Computing



OSCAR 2: Key changes

- New OSCAR Wizard based on webmin
 - ◆ Installation modes
 - Simple – Standard – Expert
- OSCAR & OS install separate
- Maintenance modes
 - Node: add – delete – update – reinstall
 - Package: add – delete
 - Configuration report

The OSCAR Solution Stack for Cluster Computing

OSCAR component packages

- OSCAR consists of a set of components built around a core OSCAR infrastructure.
- OSCAR Data Repository (ODR)
 - ◆ read – by anyone
 - ◆ write – by OSCAR wizard only
- We will publish the interfaces required to turn a software package into an OSCAR component package.
 - Add/delete package to a node/system.
 - Modify package configuration for add/delete nodes.

An OSCAR Packages API will enable a community of HPC software developers to spring up around OSCAR.

OSCAR 2 – Security Options

- Wizard based
 - ◆ Security options selected in wizard installer
- Security schemes
 - ◆ All Open
 - ◆ Nodes isolated to private subnet
 - ◆ Cluster firewall / NAT
 - ◆ Independent packet filtering per node
- Probably will use “pfilter”
<http://pfilter.sourceforge.net/>

The OSCAR Solution Stack for Cluster Computing

OSCAR Development Path

- **version 1.0**
 - ◆ Redhat 6.2 based
 - ◆ Nodes built by LUI (IBM)
 - ◆ Proof of concept (prototype)
 - ◆ Many steps, sensitive to bad input
 - ◆ Flexibility was intention; identify user needs
- **version 1.1**
 - ◆ Redhat 7.1 based
 - ◆ Nodes built by LUI
 - ◆ More automation for homogenous clusters
 - ◆ SSH: user keys instead of host keys
 - ◆ Scalability enhancements (ssh, PBS)
 - ◆ Latest software versions

OSCAR Development Path

- **version 1.2**
 - ◆ moved development to SourceForge www.sourceforge.net
 - ◆ LUI replaced by SIS
 - ◆ Redhat 7.1 based
 - ◆ Packages adjust to SIS based model
 - ◆ Latest software versions (C3 tools, PBS, MPICH)
 - ◆ Start releasing monthly
- **version 1.21 (1.3 beta?)**
 - ◆ Redhat 7.2 support
- **version 1.3**
 - ◆ Add/Delete node support implemented
 - ◆ Security configuration on head node
 - ◆ ia64 support

The OSCAR Solution Stack for Cluster Computing

OSCAR Development Path (cont.)

- **version 1.4**
 - ◆ Grouping support (nodes)
 - ◆ GUI replacement: Webmin (command line backend)
 - ◆ Core packages read/write configuration to database
 - SSH, C3, SIS, Wizard
 - ◆ Package DB API published
 - modular package support
- **version 1.5**
 - ◆ Existing packages use database
 - PBS, MPICH, PVM, LAM, Maui

OSCAR Development Path (cont.)

- **version 1.6 (2.0 beta?)**
 - ◆ custom security configuration for compute nodes
 - ◆ single head node model expires
 - head node holds OSCAR database
 - packages can designate their own head node (e.g. PBS)
 - ◆ package writing opened to community
 - ◆ the modularity advantage
 - “open packages” and “certified packages”
 - commercial packages can now be offered
 - licensing issues disappear
 - compatibility with other packagers (hopefully)

The OSCAR Solution Stack for Cluster Computing

Agenda

- Cluster Computing Software Stacks
 - What they are and why we desperately need a common base software stack.
- OSCAR and the Open Cluster Group
- OSCAR Today
- OSCAR Tomorrow
- ➡ • Cluster Computing Standards

Cluster Computing Platforms

- In an ideal world ...
 - ♦ There would be two base cluster computing platforms:
 - Symmetric Clusters (AKA traditional Beowulf)
 - Single System Image Clusters
 - ♦ Different groups would use one of these 2 base platforms and extend them to meet their needs.
- But in the real world ...
 - ♦ Everyone wants their own platform.

HPC research groups seeking funding win with multiple platforms.

Vendors, end users, and ISV suffer due to a lack of commonality.

The OSCAR Solution Stack for Cluster Computing

Cluster Computing Standards

- If we can't share base cluster computing platforms, let's at least share:
 - Interfaces for defining cluster-enabled software packages
 - File system organization – an application shouldn't have to figure out where MPI is on each different cluster
 - Data base API for cluster configuration information.

The goal should be to make cluster look the same to end users and ISV's.

Cluster Computing Standards

- Cluster Computing standards options:
 - **Global Grid Forum**
 - This would work, but it might be too slow and cumbersome.
 - **Ignore everyone else and one of us does such a good job, a de facto standard emerges**
 - This is what will hopefully happen if we can't get cluster computing groups to collaborate.
 - **The major two or three players just make it happen**
 - What if the OpenClusterGroup, PCCC and Scyld just agreed on a standard? That's more than 60% of the market. Everyone else would have to follow.

The OSCAR Solution Stack for Cluster Computing

Summary

- Software drives the HPC market.
- For HPC clusters to make it for commercial, numerically intensive computing; we need a common robust cluster Platform.
- Industry, academic and national-lab HPC professionals **MUST** work together to make this happen: