



OpenSSI Linux Clustering for HPC on Itaniums

www.openssi.org

Dr. Bruce J. Walker
HP Fellow
Office of Strategy and Technology

Many types of Clusters

High Performance Clusters

Beowulf; 1000 nodes; parallel programs; MPI

Load-leveling Clusters

Move processes around to borrow cycles

Web-Service Clusters

LVS; load-level tcp connections; replicate data

Storage Clusters

Lustre; GFS; parallel filesystems; same view of data from each node

Database Clusters

Oracle RAC;;

High Availability Clusters

ServiceGuard, Lifekeeper, Failsafe, heartbeat, failover clusters

Clustering Goals

One or more of:

High Availability

Scalability

Manageability

Usability

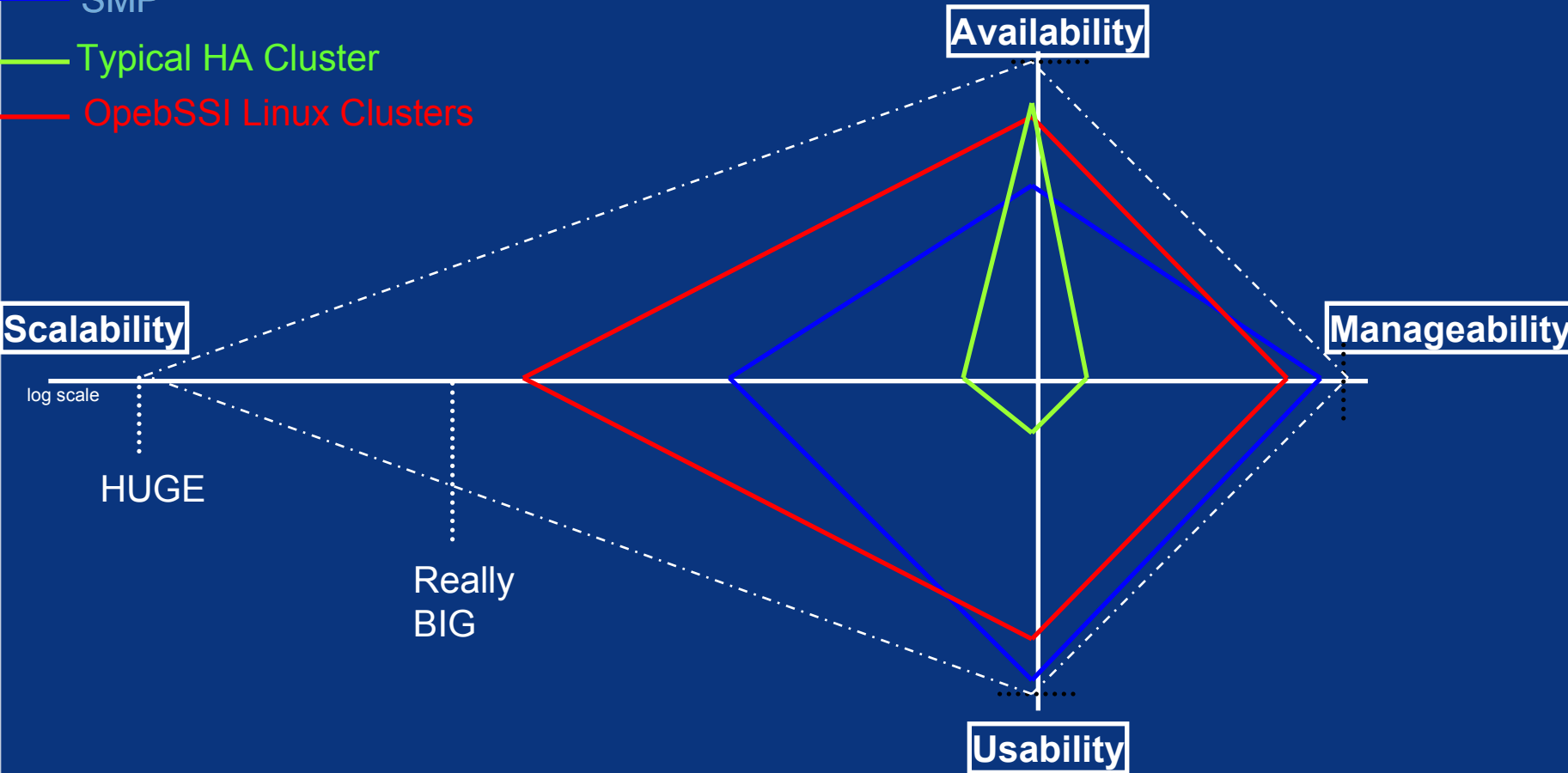
OpenSSI Linux Clusters

----- Ideal/Perfect Cluster in all dimensions

SMP

Typical HA Cluster

OpebSSI Linux Clusters



Overview of OpenSSI Cluster

- **Standard Hardware with TCP/IP between the nodes**
 - **- use Quadrics/Myrinet for MPI and/or Lustre**
- **Single HA root filesystem**
- **Consistent OS kernel on each node**
- **Join cluster early in boot**
- **Peer operation rather than master-slave**
- **Single view of filesystems, devices, processes, ipc objects**
- **Single management domain**
- **Load balancing of connections and processes**
- **Failover of services and applications**

Availability

- **No Single (or even multiple) Point(s) of Failure**
- **Automatic Failover/restart of services in the event of hardware or software failure**
- **Filesystem failover integrated and automatic**
- **Application Availability is simpler in an SSI Cluster environment; statefull restart easily done;**
- **Application transparent checkpoint/restart**
- **Architected to avoid scheduled downtime**
 - **Node eviction via transparent process migration**

Price / Performance Scalability

What is Scalability?

Environmental Scalability and Application Scalability!

Environmental (Cluster) Scalability:

- more USEABLE processors, memory, I/O, etc.
- SSI makes these added resources useable

Price / Performance Scalability - Application Scalability

- **SSI makes distributing function very easy**
- **SSI allows sharing of resources between processes on different nodes**
- **Replicated instances easily co-ordinate**
- **Monolithic applications don't "just" scale**
- **Selective Load balancing**
 - connections and processes

OpenSSI Clusters

Price/Performance Scalability

- **SSI allows any process on any processor**
 - **general load leveling and incremental growth**
- **All resources transparently visible from all nodes:**
 - **filesystems, IPC, processes, devices*, networking***
- **OS version in local memory on each node**
- **Migrated processes use local resources and not home-node resources**
- **Industry Standard Hardware (can mix hardware)**
- **OS to OS messages minimized**
- **Distributed OS algorithms written to scale to hundreds of nodes (and successful demonstrated to 133 blades and 27 Itanium SMP nodes)**

OpenSSI Linux Clusters

What about Manageability and Ease-of-Use?

SMPs are easy to manage and easy to use.

SSI is the key to manageability and ease-of-use for clusters

OpenSSI Linux Clusters - Manageability

- **Single Installation**
- **Joining the cluster is automatic as part of booting and doesn't have to be managed**
- **Trivial online addition of new nodes**
- **Use standard single node tools (SSI Admin)**
- **Visibility of all resources of all nodes from any node**
- **Applications, utilities, programmers, users and administrators often needn't be aware of the SSI Cluster**
- **Simpler HA (high availability) management**

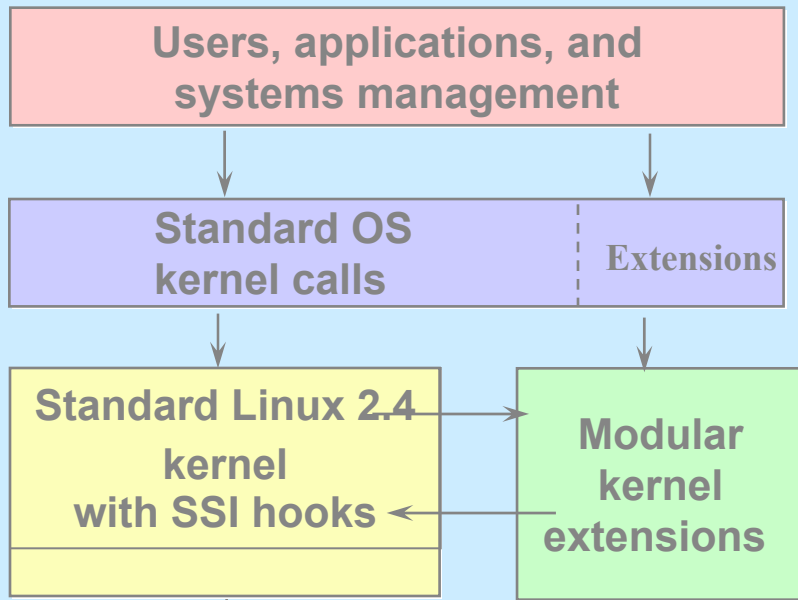
OpenSSI Linux Cluster

Ease of Use

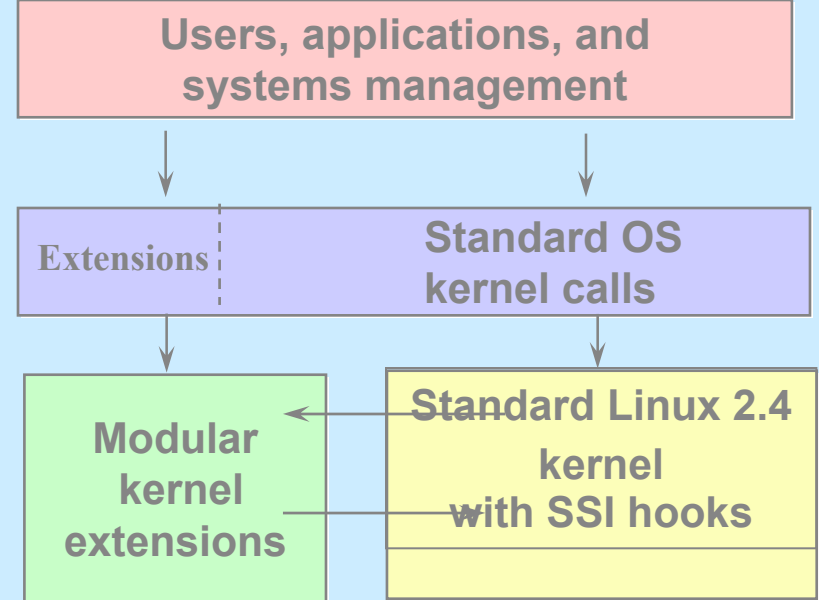
- **Can run anything anywhere with no setup;**
- **Can see everything from any node;**
- **Service failover/restart is trivial;**
- **Automatic or manual load balancing;**
- **Powerful environment for application service provisioning, monitoring and re-arranging as needed**

How Does SSI Clustering Work?

Uniprocessor or SMP node

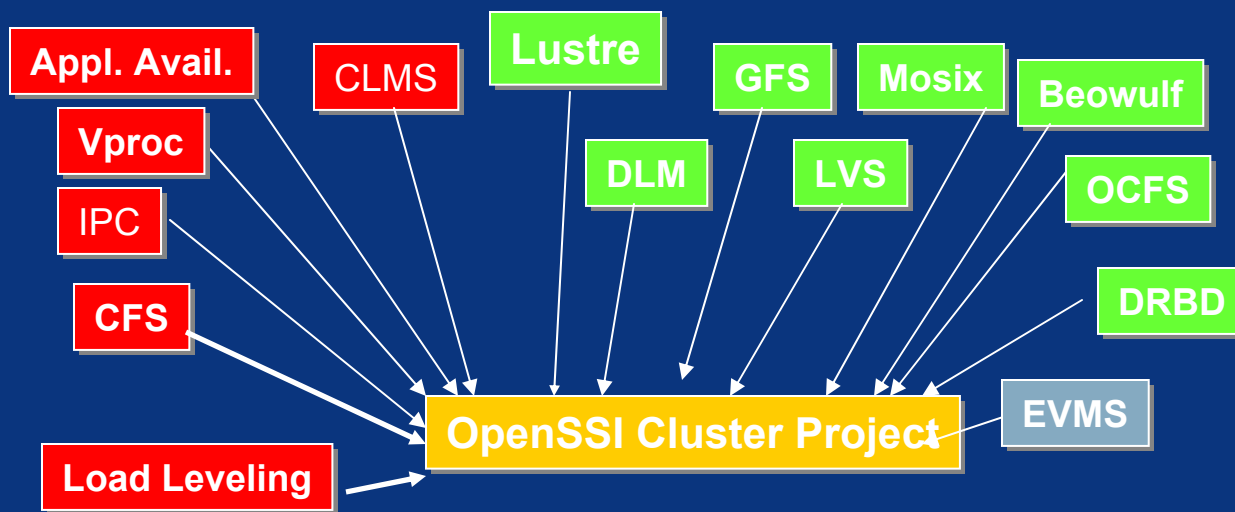


Uniprocessor or SMP node



Other nodes

Component Contributions to OpenSSI Cluster Project



 HP contributed

 Open source and integrated

 To be integrated

Process Model in OpenSSI - Vproc

Single clusterwide pid space

Full /proc for all processes from all nodes

Full function rexec() and rfork() and process migration

Distributed process relationships

Full job control, signalling, inheritance

Moved processes execute fully on the new node

Full handling of node failures

Not master-slave oriented; full peer orientation

Transparent Load leveling of processes at exec time and while they are running;

Checkpoint/restart

Load Leveling in OpenSSI

Have connection load balancing via LVS;

Have process load balancing via load leveler:

- can selectively turn it on (on a node basis)
- can select which applications to load balance
 - attribute inherited so specifying “make” will load balance the cc’s, etc.
- transparent balancing at exec() time if possible;

Component Contributions to OpenSSI Cluster Project

Lustre:

- open source project, funded by HP, Intel and US Labs
- parallel network filesystem;
- file service split between a metadata service (directories and file information) and data service (spread across many data servers (stripping, etc.))
- OpenSSI has Lustre client support integrated

Component Contributions to OpenSSI Cluster Project

LVS - Linux Virtual Server:

- front end director (software) load levels connections to backend servers; director is part of OpenSSI cluster
- can use NAT, tunneling or redirection
 - (we are using redirection)
- can failover director (HA-LVS capability)
- integrated with OpenSSI membership;

NFS in OpenSSI

Cluster as an NFS client:

- any node does a mount and automatically mounted on all nodes;
- each node goes direct to the NFS server;
- full file locking and nodedown handling;

Cluster as an NFS server:

- mounts come in thru the LVS CVIP;
- have HA-NFS except for file locking;

Interconnects and OpenSSI

OpenSSI protocols run over tcp/ip and will typically use 100Mb or 1gigabit;

If you have Myrinet or Quadrics, the MPI and/or Lustre traffic can go over those interconnects (have tested Myrinet but not Quadrics yet).

An activity to get native Infiniband support should start this summer

HPTC Middleware

Have a package (currently IA-32) of:

MPICH (modified and not), ScalablePBS, Maui,
SLURM, Ganglia, Supermon, LAMPI (modified)

Working on Itanium versions

Have also run HP MPI and TotalView

Checkpoint/Restart

In the lab we have a modified kernel and a modified HP MPI that can be used to do transparent checkpoints of MPI applications running over TCP/IP (haven't tested on Myrinet).

Checkpoint/Restart can also be used to checkpoint simple processes.

OpenSSI Linux Clusters - Status

Version 1.0 just released –

Binary, Source and CVS options

Functionally complete RH9 and RHel3

Debian release also available

IA-32 and Itanium (RHel3 only) Platforms

Runs HPTC apps as well as Oracle RAC

Available at OpenSSI.org

**A Dozen people inside HP working full
time on this project**

Where to get Itanium version of OpenSSI

<http://ci-linux.sourceforge.net/rhel/openssi-rhel3-1.0.0-rc5.ia64.tar>

RHel3, update 2

OpenSSI Linux Clusters - Conclusions

- HP has recognized that Linux clusters are important part of the future.
- HP has recognized that combining scalability with availability and manageability/ease-of-use is key to clustering
- HP is leveraging its merger with Compaq (Tandem/Digital) to bring the very best of clustering to a Linux base
- We are anxious to get Itanium users of OpenSSI and will support those doing so