51CTO.com 技术成就梦想 | UBM

2012 / 10 / 25 ► 2012 / 10 / 26
中国 · 北京
October 25-26,2012
Beijing · China

World of Tech WOT

2012云计算架构师峰会

Cloud Computing Architects Summit China 2012

揭示企业级IT架构转型 分享最新技术的应用落地

主办单位：51CTO.com 技术成就梦想　联合主办：UBM

高可用集群的开源解决方案的发展与实现

北京酷锐达信息技术有限公司

技术总监 史应生

shiys@solutionware.com.cn

You may know these words

- OCF
- SPOF
- Heartbeat
- OpenAIS
- Pacemaker
- RHCS
- Keepalive
- Roseha
- Quorum
- Qdisk
- Storage Mirroring
- Multipathing
- DRBD
- GFS
- CLVM
- Fencing
- Power Switch
- Fabric Switch
- STONITH
- Bonding
- multicast
- broadcast
- Failover Domains
- Resource Manager
- CRM
- luci
- ricci
- system-config-cluster
- Yast
- Corosync
- Cman
- Heuristic
- CCS
- Virtual Synchrony
- Totem
- clusvcadm

Availability

- Mean Time Between Faults (MTBF)
  - MTBF measures the reliability of the system components
  - MTBF decreases as the number of components in the system increases
- Mean Time to Repair (MTTR)
  - MTTR measures the time to restore service
  - MTTR is decreased by the use of redundant components that allow service to continue or to be restored, even though the faulty component has not yet been repaired
- Availability – Probability that when a service is requested, it will be provided
  - If the time required to satisfy each service request is short

$$\text{Availability} = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}$$

MTBF/MTTR

- Problem
  - Increase MTBF to very large values
  - A real system has a large number of hardware and software components; thus, it is hard to achieve high availability by increasing the MTBF
- Solution
  - Reduce MTTR to very low values
  - Use redundant components to achieve high availability by decreasing the MTTR to close to 0, from hours or minutes to a few milliseconds

High Availability

- Availability
  - A measure of the uptime of a system
  - The probability that the system is able to provide service when requested
- High Availability
  - At least 99.999 %
  - A good HA clustering syste m adds a "9" to your base availability
  - 99->99.9, 99.9->99.99, 99.99->99.999, etc.
  - It cannot achieve 100% availability - nothing can

Availability	Downtime per Year
90 %	36.5 days
99 %	3.5 days
99.9 %	9 hours
99.99 %	52 min
99.999 %	5 min
99.9999 %	30 sec
99.99999%	3 sec

High Service Availability

+

=

System is available for use 99.999% of the time or more

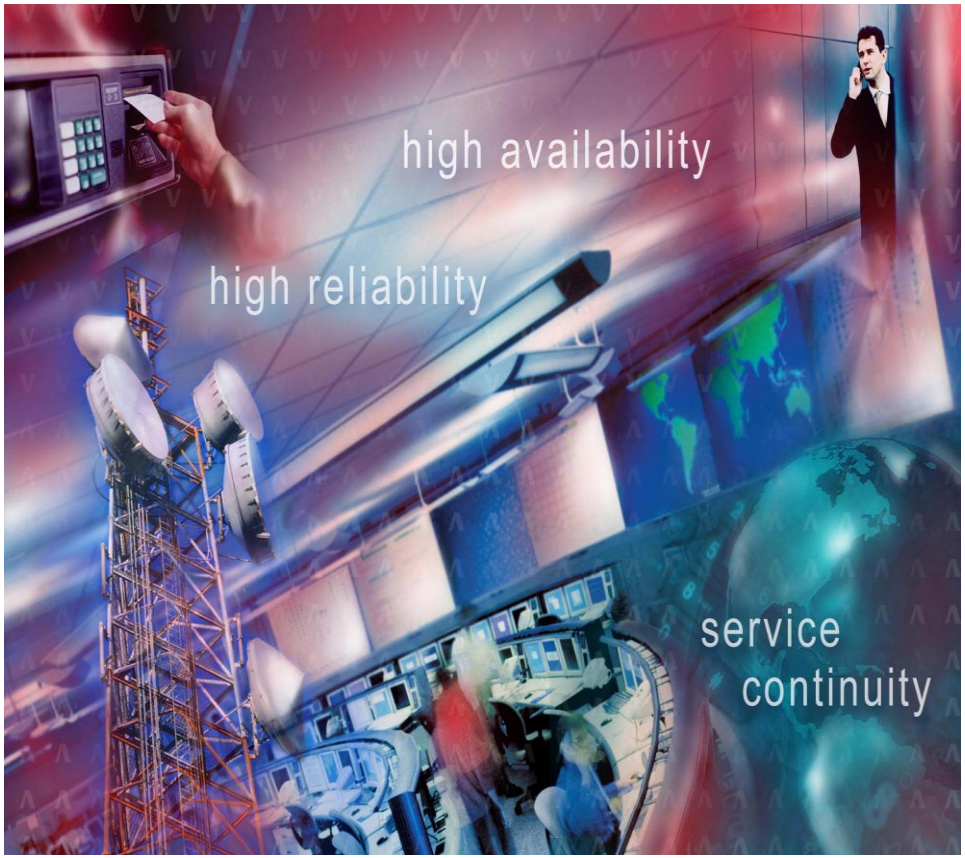
No loss of service continuity during fault recovery or administrative actions

What end-users expect

SA Forum-Open Specifications for Service Availability

SERVICE AVAILABILITY™ FORUM

Open Specifications for Service Availability

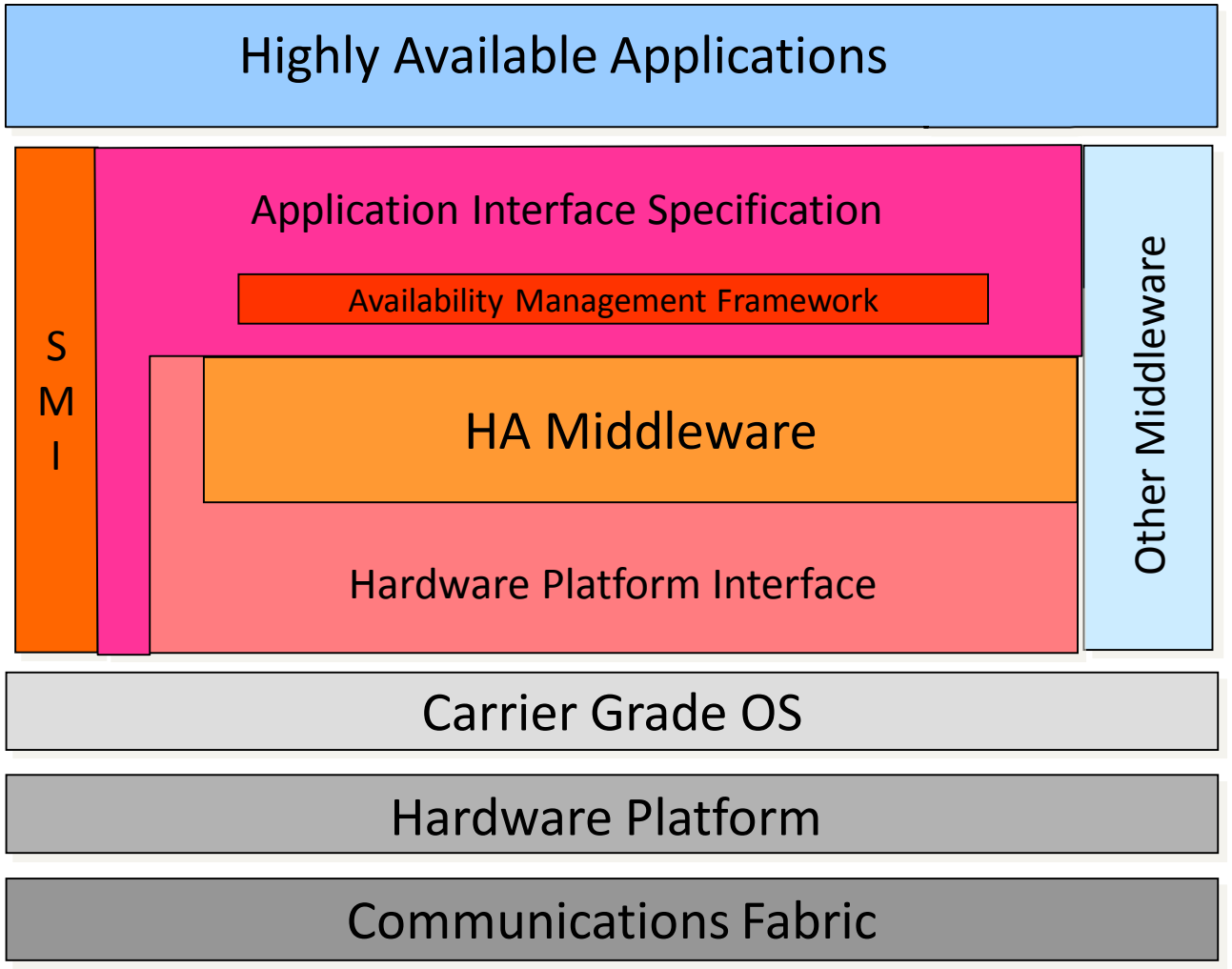


Service Availability Forum - The Software Stack

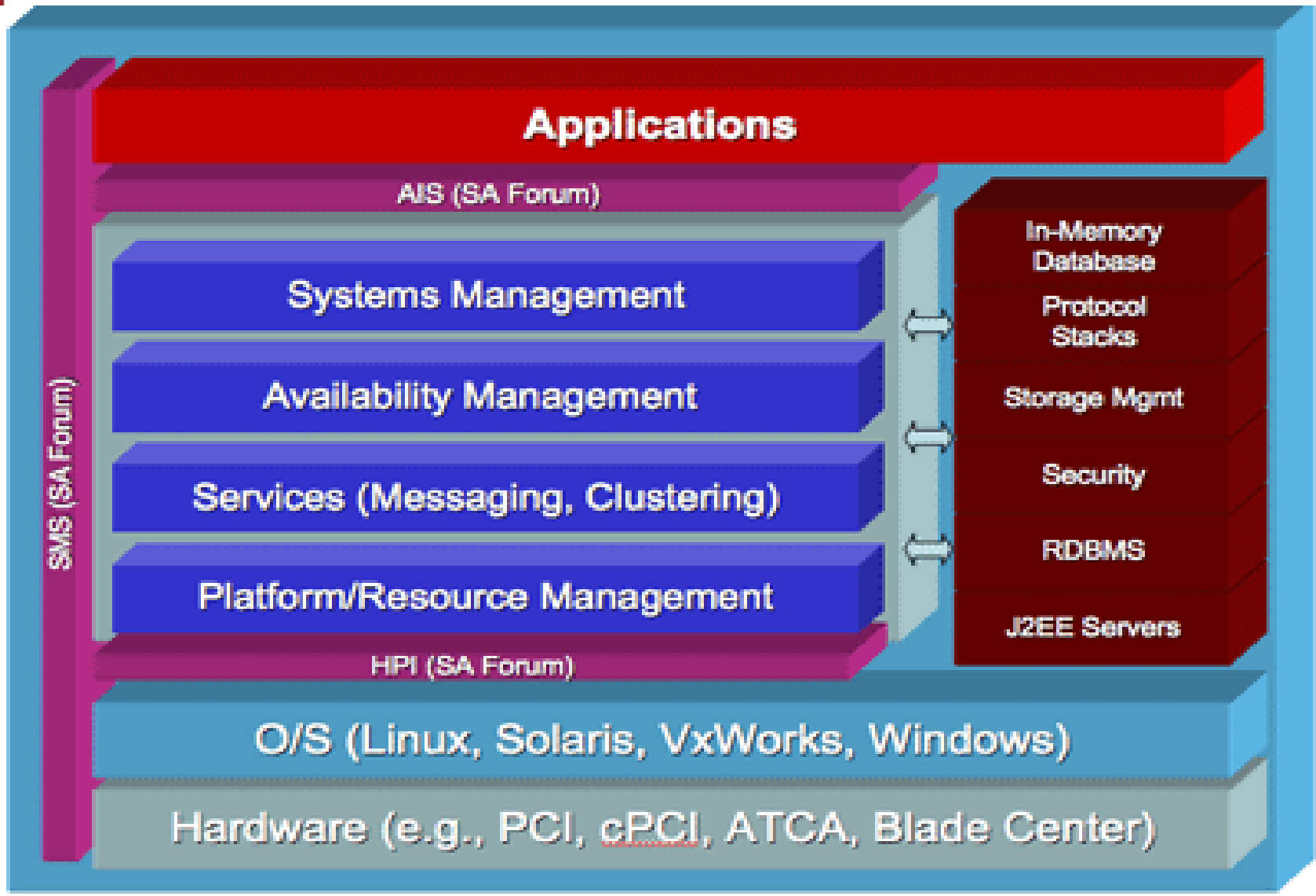
Systems Management Interface (SMI)

Application Interface Specification (AIS)

Hardware Platform Interface (HPI)



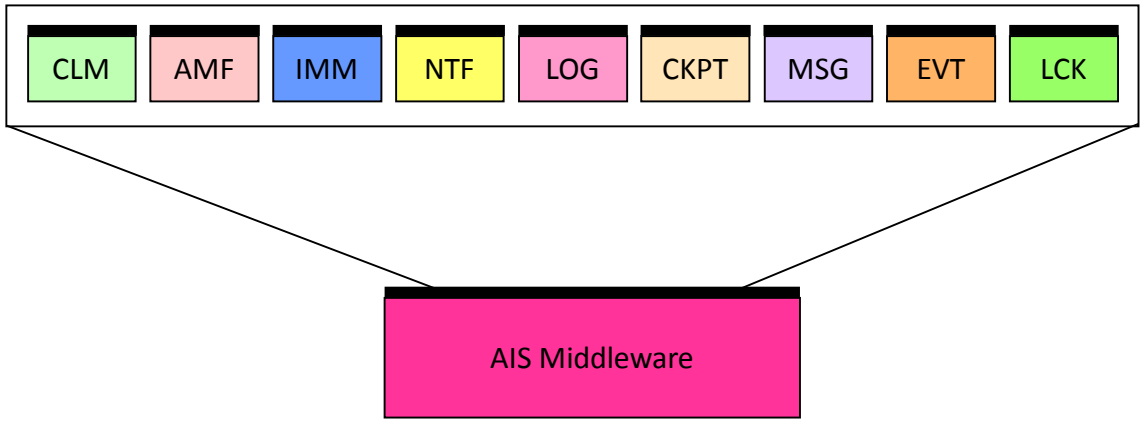
Service Availability Forum - The Software Stack

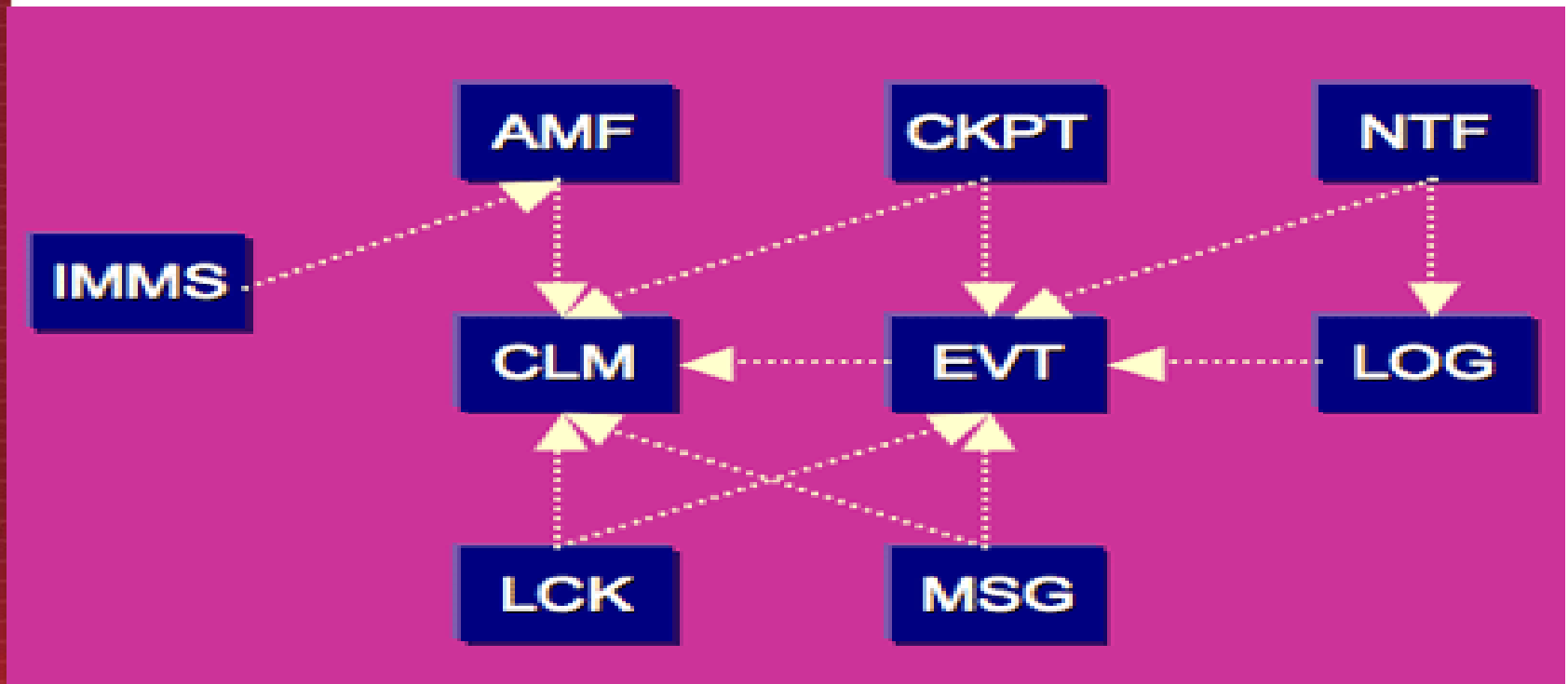


Application Interface Specification

- The AIS defines an Application Program Interface (API) for middleware between the applications and the operating system
- The AIS is divided into the following parts or areas:
  - Availability Management Framework
  - Cluster Membership Service
  - Checkpoint Service
  - Event Service
  - Message Service
  - Lock Service
  - Notification Service
  - Log Service
  - Information Model Management Service

AIS Middleware





Openais

- The main developers of this project have decided not to continue further development of the AIS implementation.
- Instead, we are spending our time maintaining a great roadmap and maintenance model around Corosync for bare metal clusters.
- The developers believe Pacemaker coupled with Corosync do a great job of providing bare metal high availability. Because Corosync and Pacemaker are shipped everywhere and are widely deployed, the need for AIS services is limited.

Corosync Project

- Started life as “openais.org” in 2002
- Corosync was created from a derivative work of the openais project
- Announced Corosync in July 2008, First 1.0.0 release in July 2009
- The Corosync Cluster Engine is a Group Communication System with additional features for implementing high availability within applications.
- Project Philosophy: Allow developers to create HA apps however they desire.

Openais vs Corosync

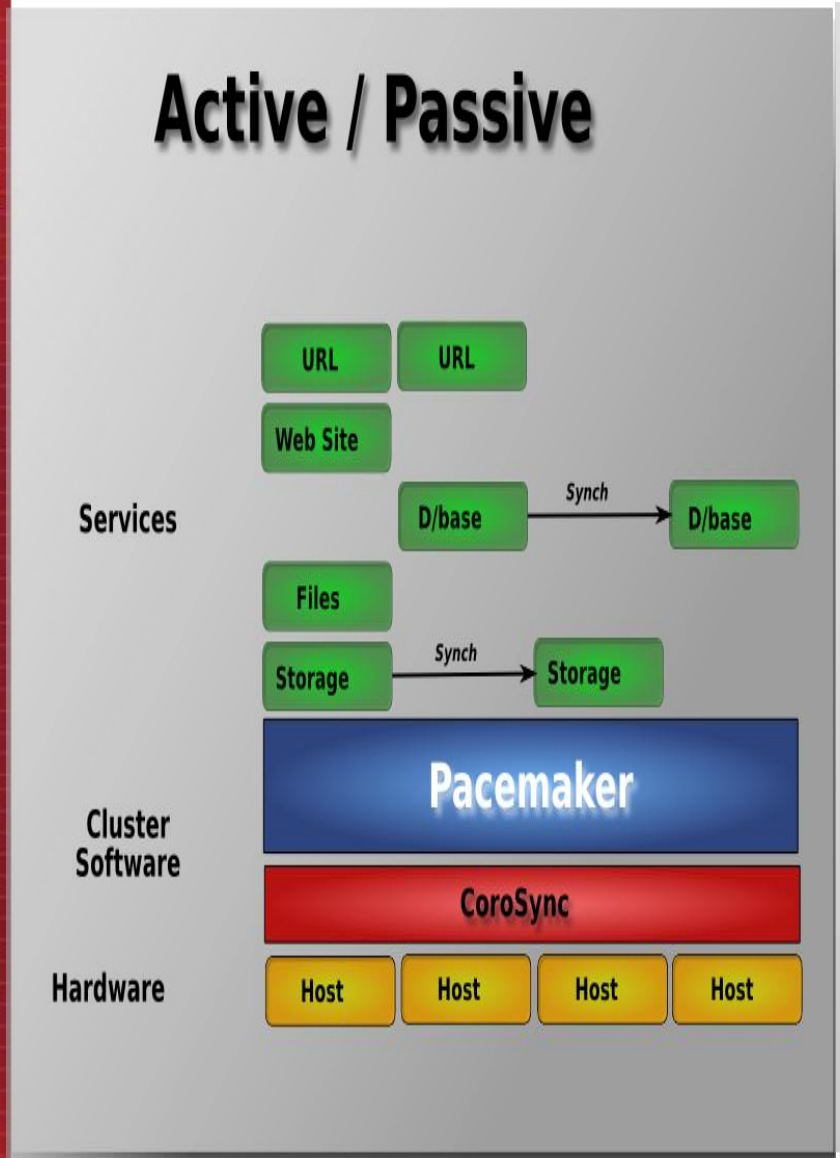
Project	branch	version	infrastructure
OpenAIS	whitetank	0.80.6	YES
OpenAIS	wilson	1.x	NO
Corosync	flatiron	1.x	YES
Corosync	needle	2.x	YES

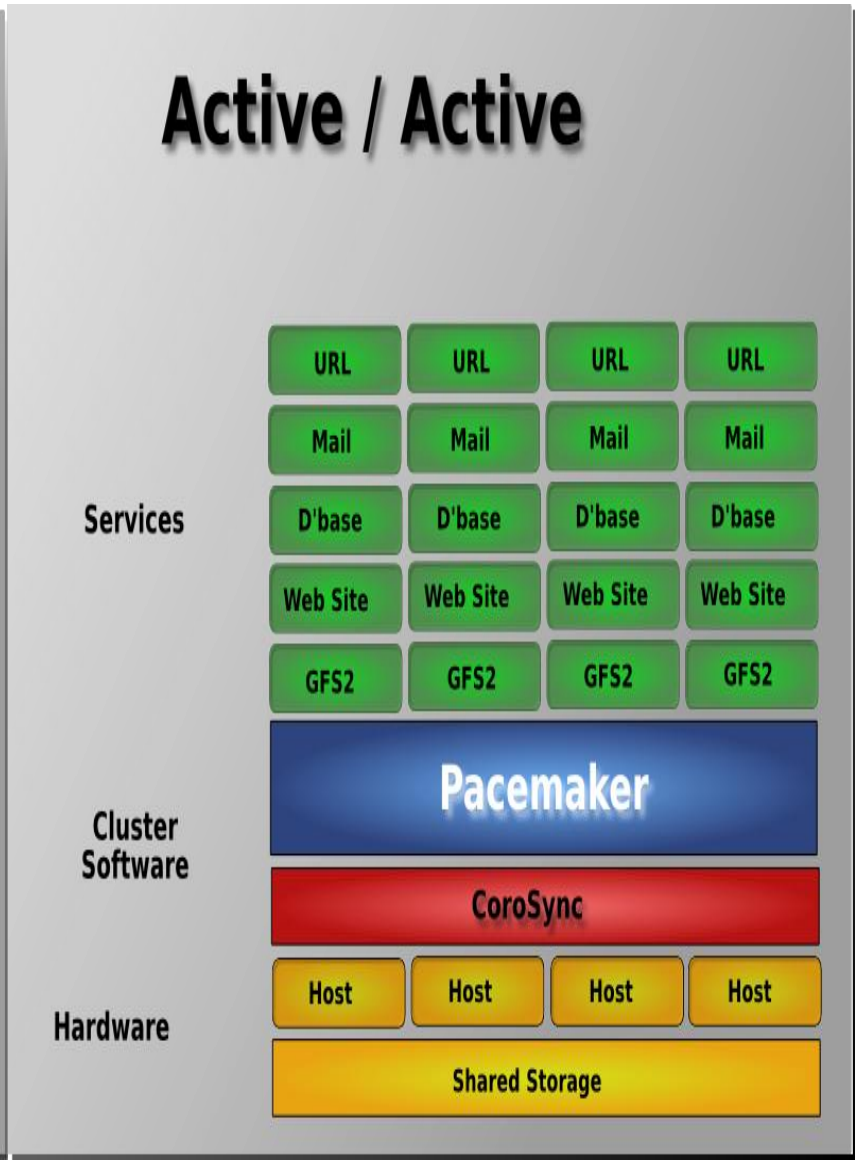
- OpenAIS whitetank was essentially split into two projects. Infrastructure went to Corosync while SA Forum APIs went to OpenAIS wilson.
- Corosync 2.x focus on small, well tested set of core services and stopped development of OpenAIS completely.

Pacemaker

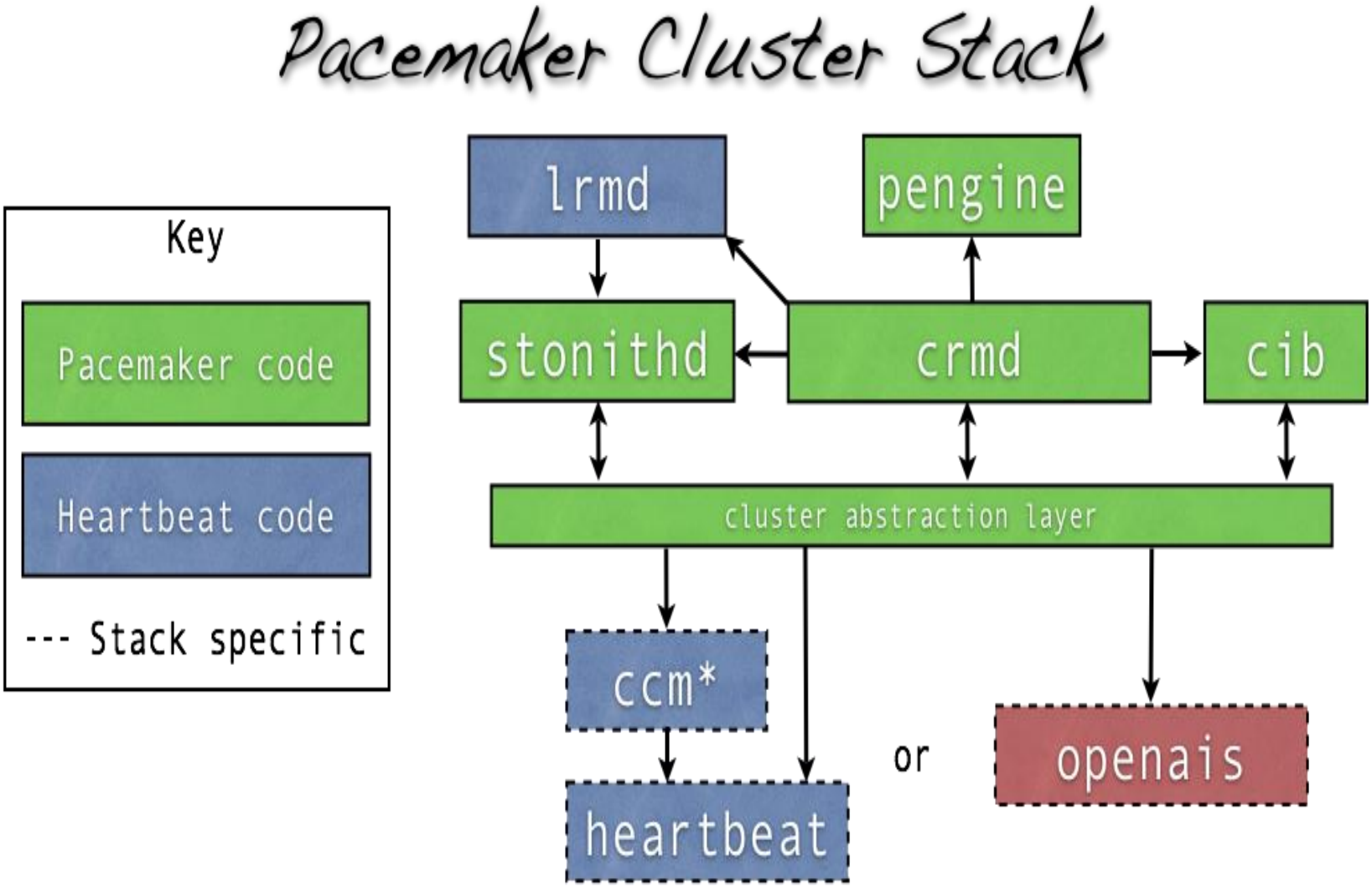
- Pacemaker is an Open Source, High Availability resource manager suitable for non-ha-aware applications for both small and large clusters.
- Pacemaker is primarily a collaborative effort between Red Hat and Novell.

Pacemaker Deployment Example

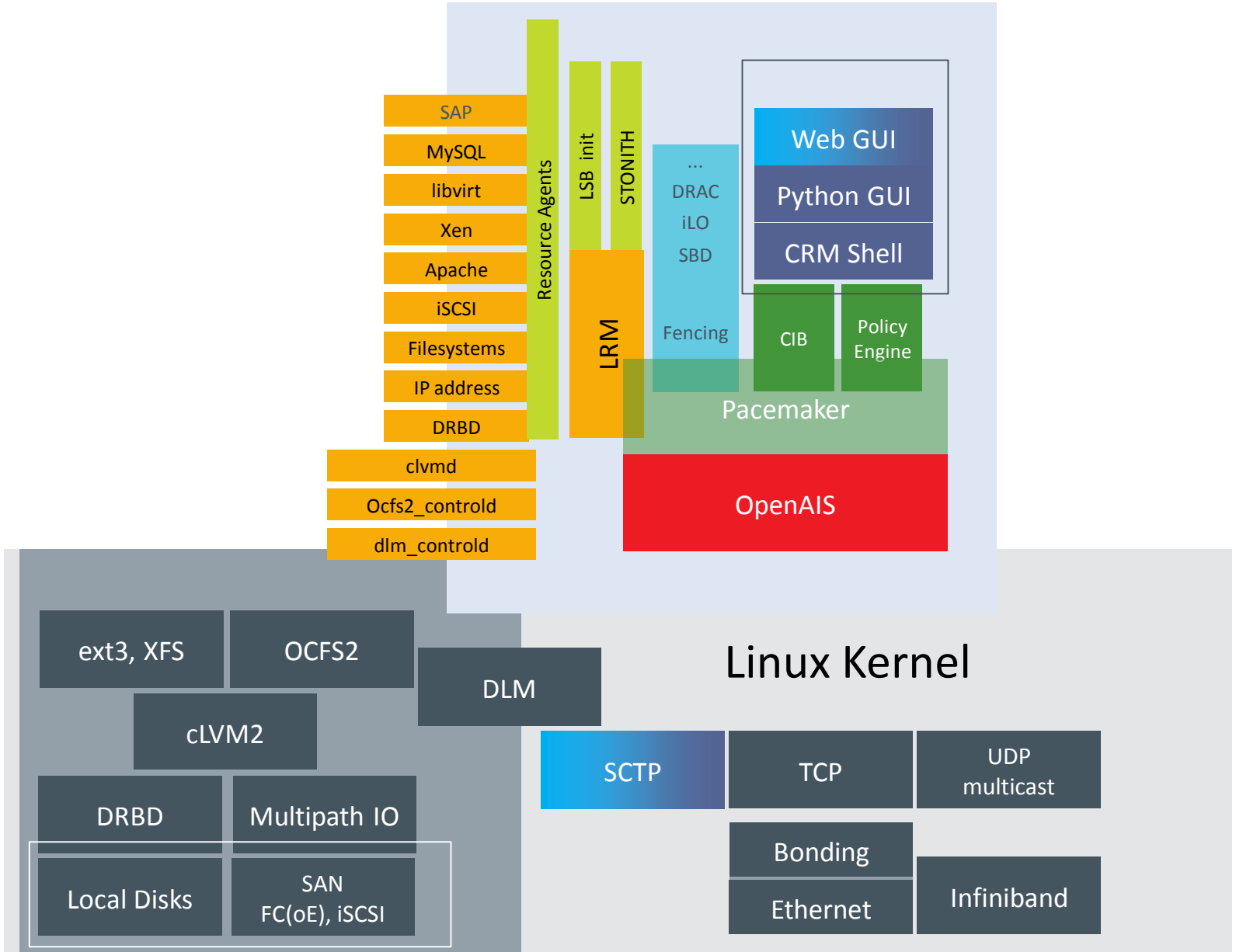




Pacemaker Cluster Stack

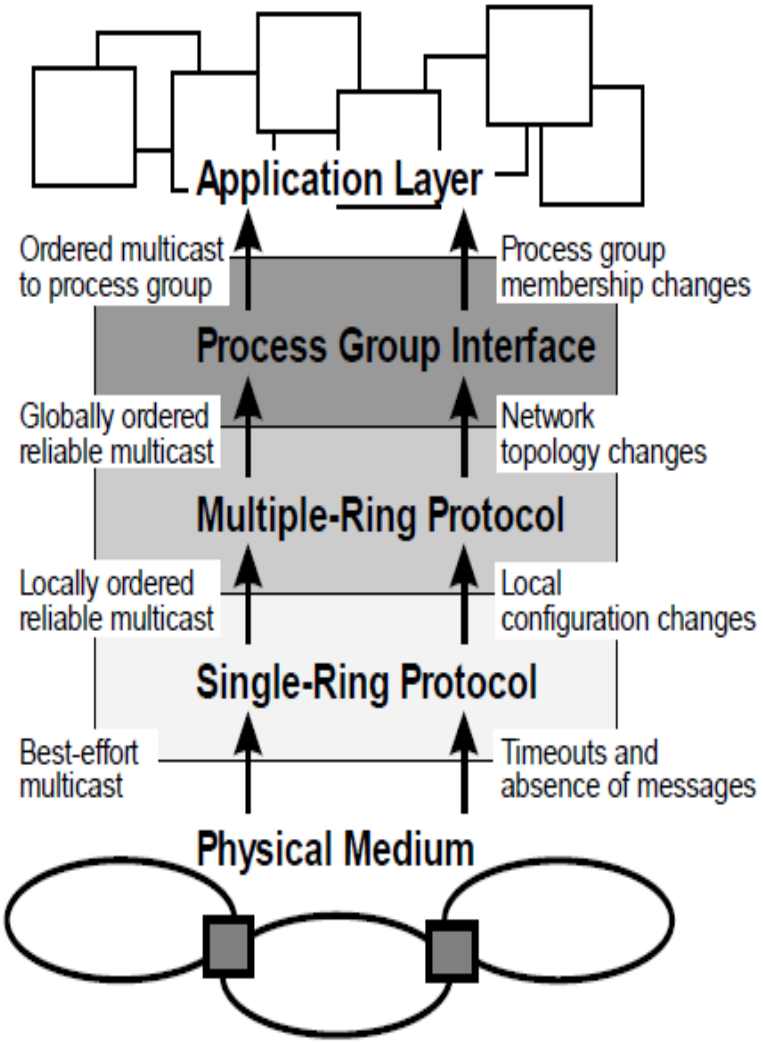


Pacemaker Cluster Stack



Totem

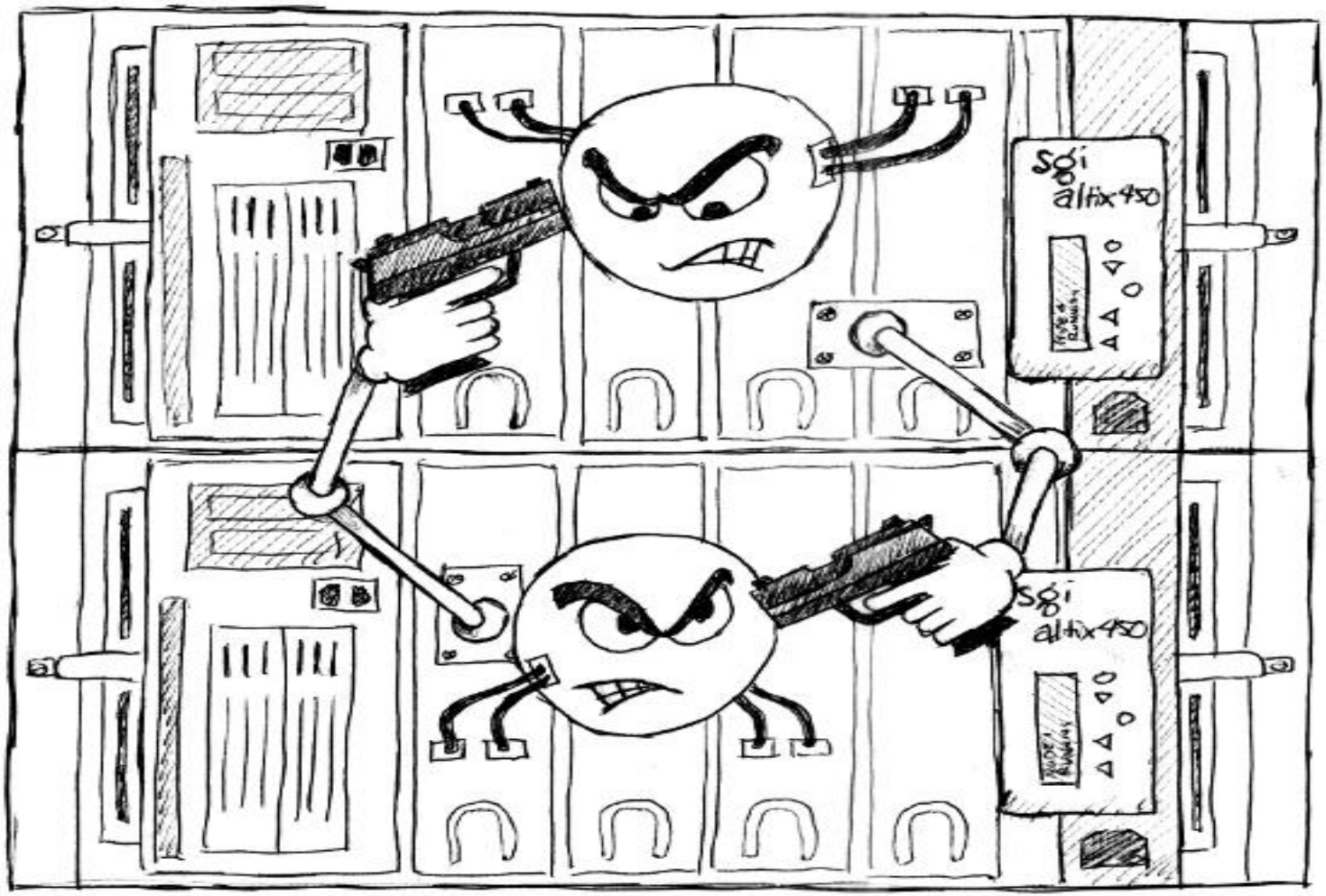
- A Fault-Tolerant Multicast Group Communication System
- High throughput and low predictable latency
- Rapid detection of, and recovery from, faults
- System-wide total ordering of messages
- Scalability to larger systems based on multiple LANs



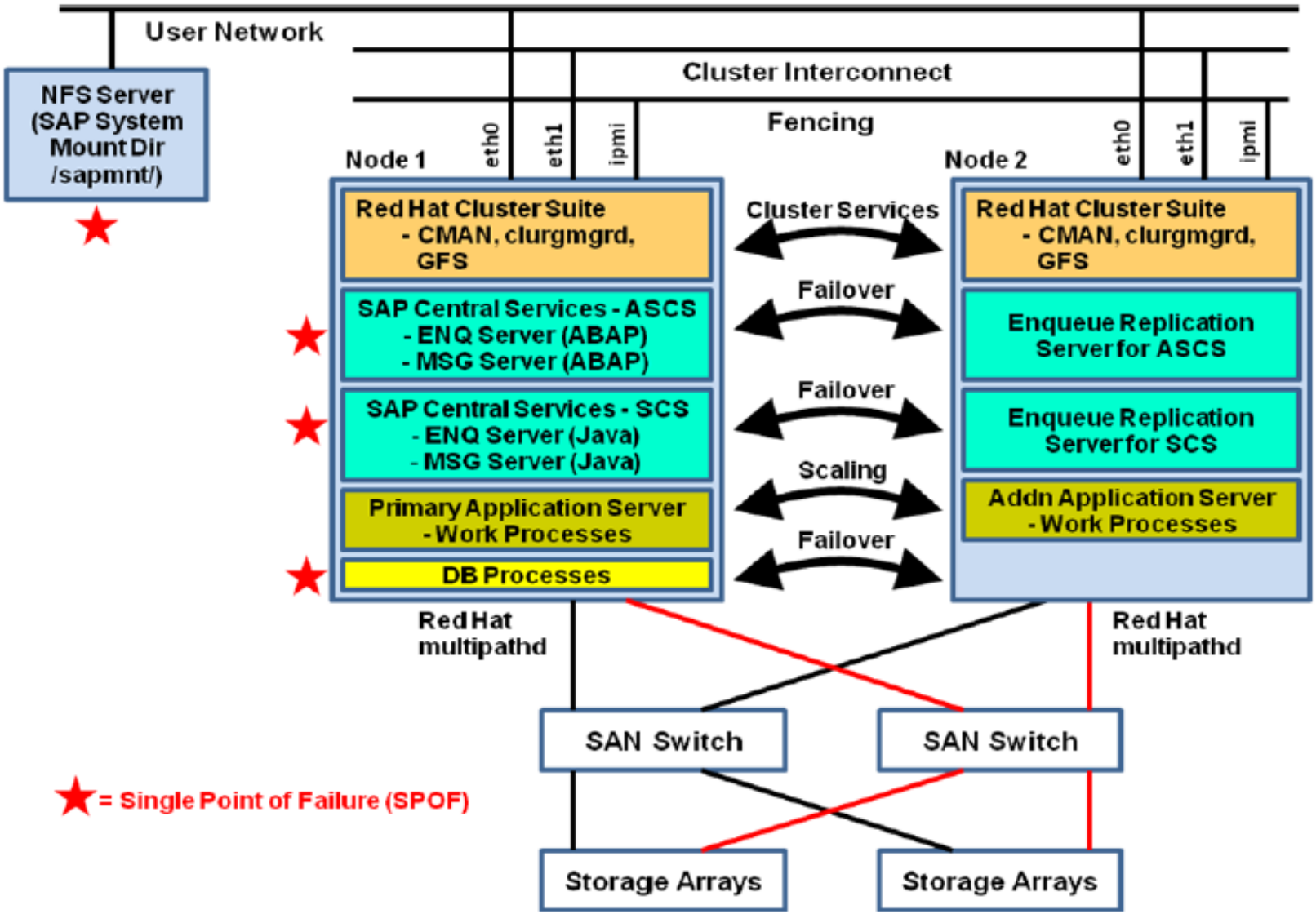
Fencing

- Fencing is a absolutely critical part of clustering. Without fully working fence devices, your cluster will fail.
- Sorry, I promise that this will be the only time that I speak so strongly. Fencing really is critical, and explaining the need for fenc
- IO Fence vs Power Fence

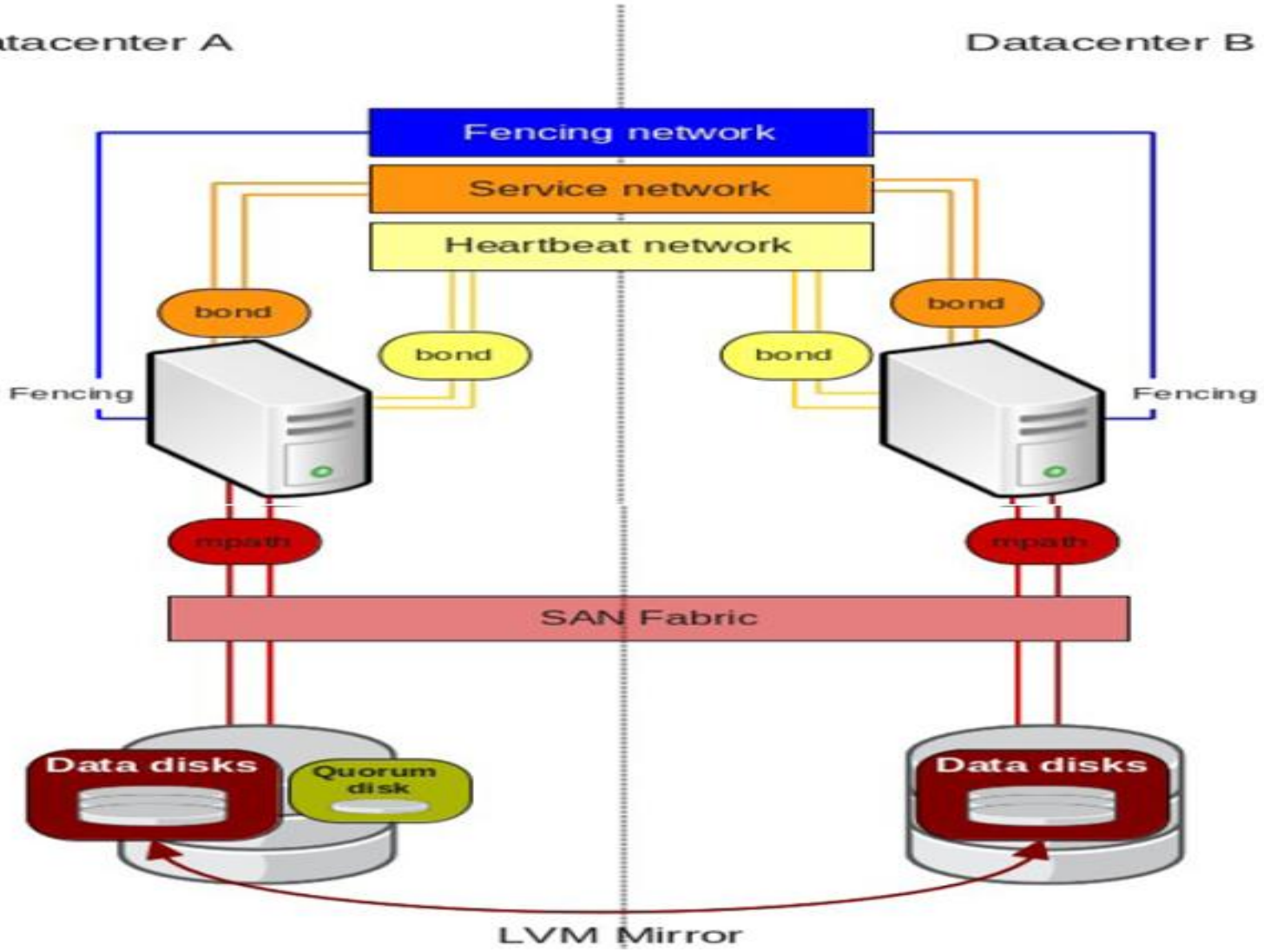
2-Node Pitfalls: Fence Loops/Fence Death



SAP Cluster



Cross city cluster



Three Sysadmin Rules

1) Backup Everything (and validate the backup regularly)

2) Master the Command Line (and avoid the UI if possible)

3) Automate Everything (and become lazy)

- Lazy sysadmin is the best sysadmin.

Who Are We?

中国领先的Linux全面解决方案提供商

Join Us . . . . . .

hr@solutionware.com.cn

酷锐达 SolutionWare

Q&A