# Pacemaker

A Scalable
Cluster Resource Manager
for Highly Available Services

Owen Le Blanc
I T Services
University of Manchester

# C V

- 1980, U of Manchester since 1985
- CAI, CDC Cyber 170/730, Prime 9955
- HP 300, HP 700, Xenix, FreeBSD, NetBSD, AIX, IRIX, Minix
- Linux 1991, mostly Debian

# Outline

- What is pacemaker?
- How do you get it?
- How does it work?
- Where did it come from?
- Documentation
- How do you configure it?
- How do you manage it?
- Real examples

# What is Pacemaker?

- If a cluster is a group of machines of any size, and
- If a resource is any application or process that can be started by a script,
- Then pacemaker allows you to manage that resource across that cluster
- in a way that makes high availability possible

# Examples

- IPVS Load Balancer
- DRBD network disk
- DB server
- Web server

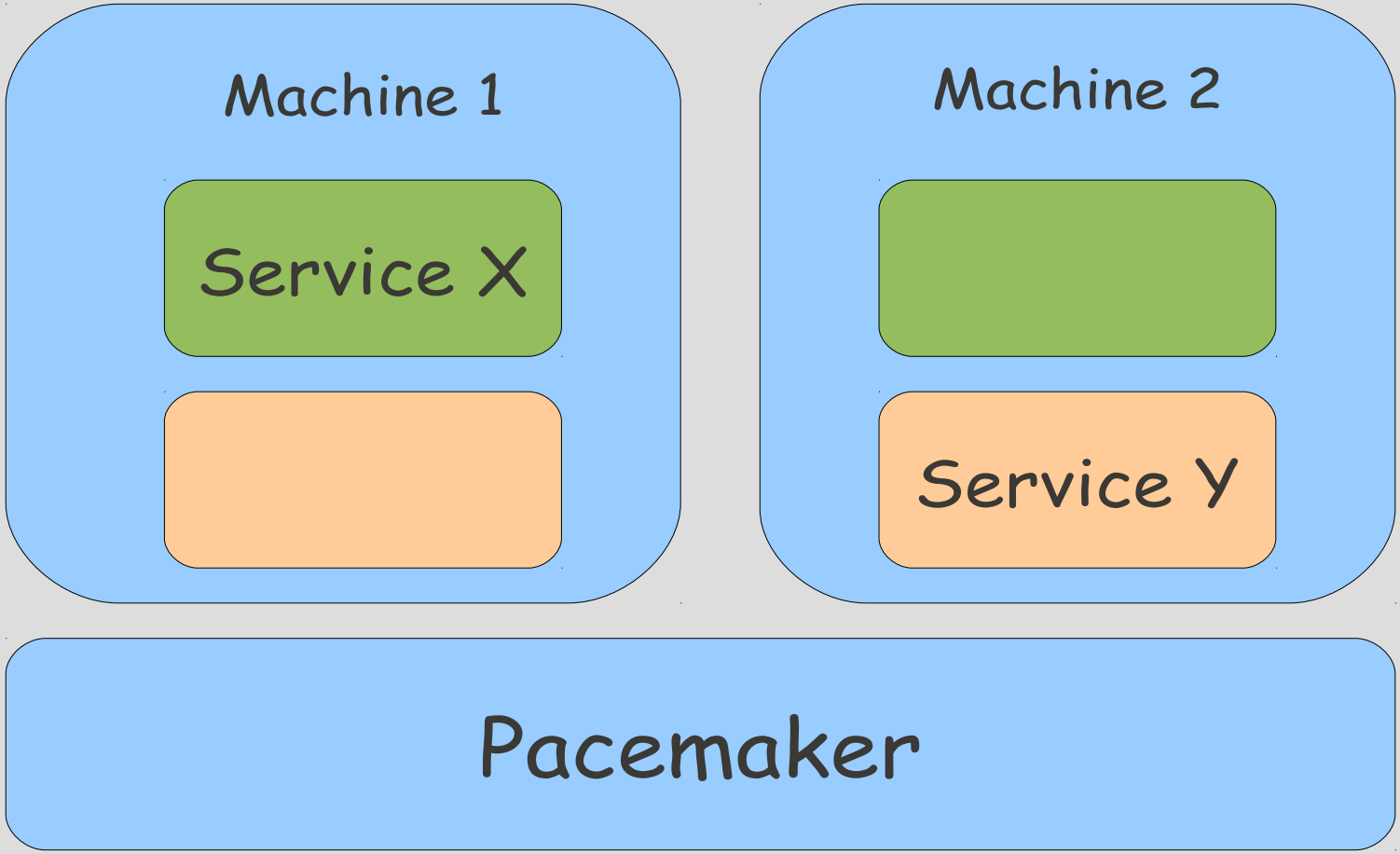
# Features

- Does not require particular storage type.
- Quorate and resource-driven clusters
- Supports various types of redundancy:
  - Active/Passive, Active/Active, Master/Slave, Multiple
- Can manage any resource controlled by shell scripts.
- Supports applications which must run on multiple machines.

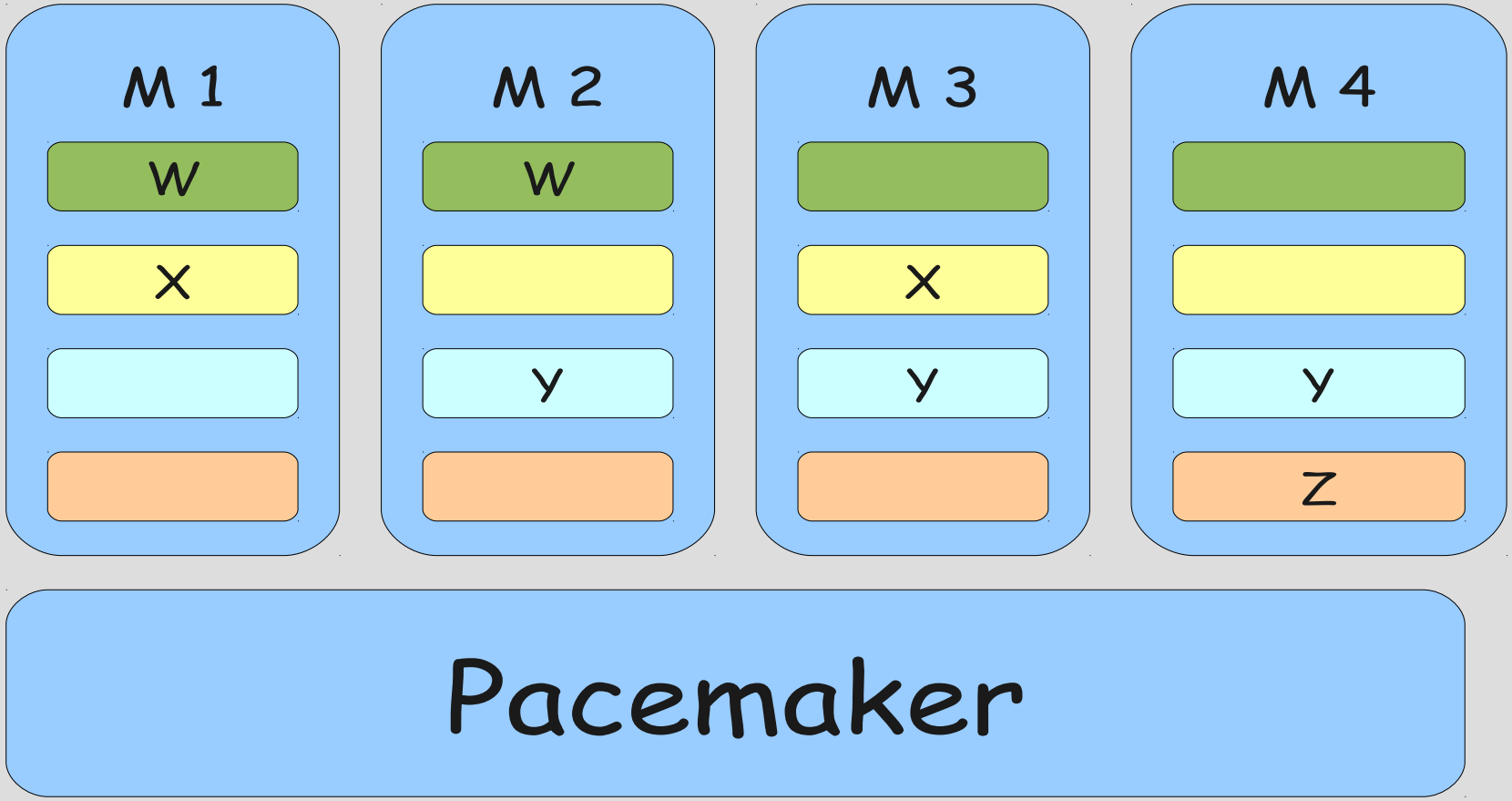
# More Features

- Detects and recovers from machine and application failures.
- Scriptable shell (command line)
- Graphical Interface(s)
- Fencing
- Ordering, location constraints

# Active/Passive

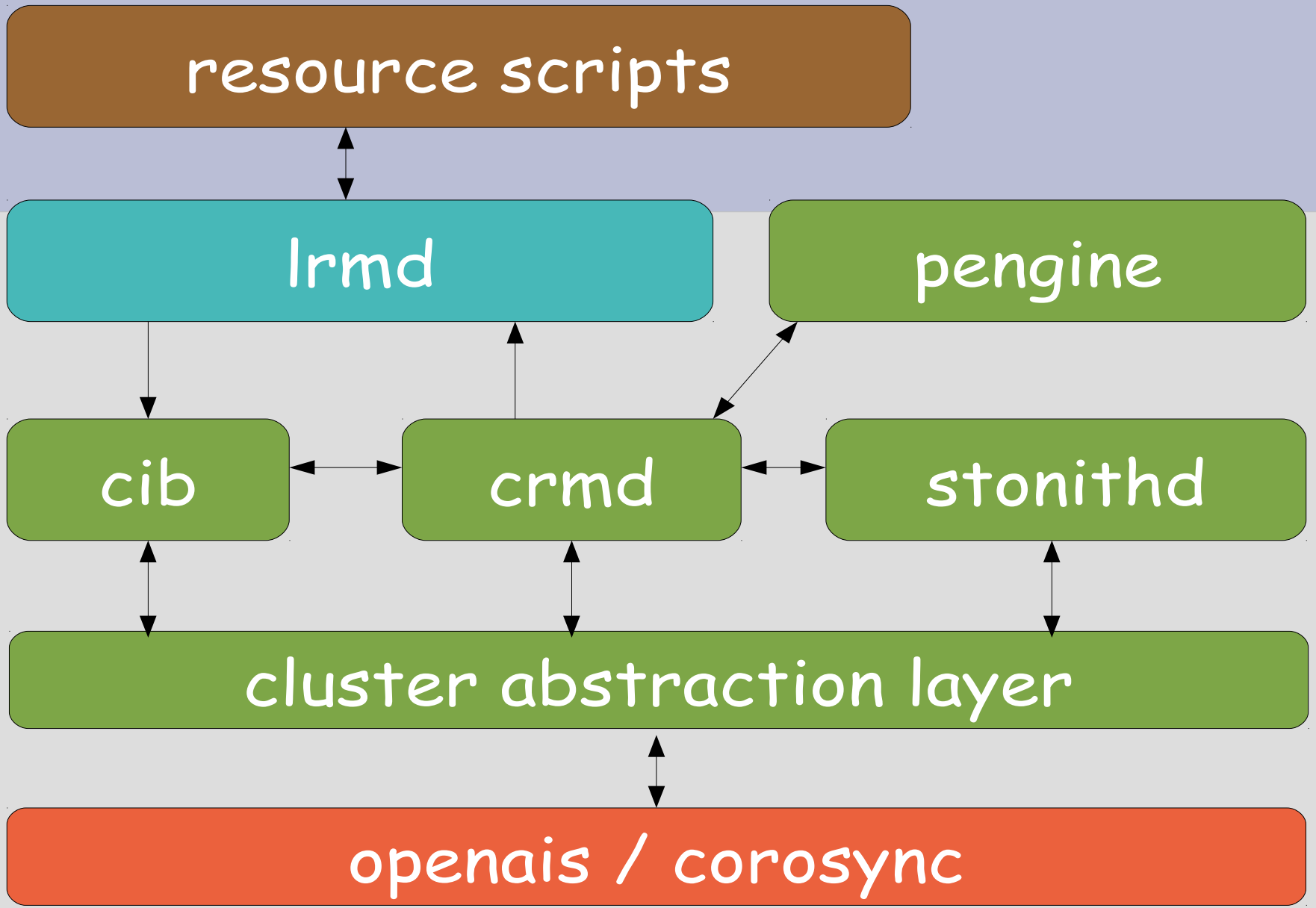


# Multiple Services



# Distributions

- Fedora, RHEL, Centos
- OpenSUSE, SLES, SLES HA
- Debian, Ubuntu
- Build from source (BSD, MacOS X, etc.)


resource scripts
lrmd
pengine
cib
crmd
stonithd
cluster abstraction layer
openais / corosync

# Internals, explained

- LMRD – Local Resource Mgmt Daemon
- CIB – Cluster Information Base
- CRMD – Cluster Resource Mgmt Daemon
- Pengine – Policy Engine
- STONITHD – Fencing subsystem
- Corosync – Communications, including cluster membership

# Parts - History

- Heartbeat - Old Linux-HA cluster manager
  - Alan Robertson, 1998 - 2007
- OpenAIS - Application Interface Spec
  - 2005 - ongoing; Corosync Cluster Engine
- Pacemaker - Cluster Resource Manager
  - Lars Marowsky-Brée, 2003 - ongoing
- Other bits
  - E.g., DRBD support from LINBIT
    - Philipp Reisner, Lars Ellenberg, 1998 - ongoing

# Old Heartbeat

- Maximum of 2 nodes
- Highly coupled design and implementation
- Simplistic group-based resource model
- Inability to detect and recover from resource-level failures

# New Pacemaker

- Much more configurable
- Much more flexible
- Much more powerful
- Much easier to control
- You can update pacemaker without interrupting the managed service(s)

# Documentation

- www.clusterlabs.org/wiki
- Clusterbau: Hochverfügbarkeit mit pacemaker, OpenAIS, heartbeat und LVS
  by Michael Schwartzkopff, O'Reilly
- A Practical Guide to XEN High Availability
  by Sander von Vugt
- www.drbd.org/users-guide/

# Configuration

- CRM shell -
  - Command line
  - File based configuration
- GUI -
  - Easier use, syntax
  - Not suited to copying whole configuration

# Example Configuration

```
primitive res-Apache ocf:heartbeat:apache \
meta is-managed="true" \
operations $id="res-apache-operations" \
op monitor interval="10" timeout="20s" \
params \
configfile="/etc/apache2/apache2.conf" \
httpd="/usr/sbin/apache2"
```

# Mysql Cluster

| Name | Status | Details |
|---|---|---|
| Cluster | have quorum | Openais & Pacemaker |
| pannier | online | |
| punnet | online (dc) | |
| Resources | | |
| grp-Ip-Fs-Msyql | group | |
| res-Wmysql3-IP | running on ['punnet'] | ocf::heartbeat:IPaddr2 |
| res-Wmysql3-Fs | running on ['punnet'] | ocf::heartbeat:Filesystem |
| res-Mysql | running on ['punnet'] | ocf::heartbeat:mysql |
| ms-Drbd-R0 | master | |
| res-Drbd-R0:0 | running (Slave) on ['pannier'] | ocf::linbit:drbd |
| res-Drbd-R0:1 | running (Master) on ['punnet'] | ocf::linbit:drbd |
| clo-Ping | clone | |
| res-Ping:0 | running on ['punnet'] | ocf::pacemaker:ping |
| res-Ping:1 | running on ['pannier'] | ocf::pacemaker:ping |

# crm status (1)

Last updated: Fri Jan 21 11:25:59 2011
Stack: openais
Current DC: punnet - partition with quorum
Version: 1.0.9-unknown
2 Nodes configured, 2 expected votes
3 Resources configured.

# crm status (2)

```
Online: [ pannier punnet ]

 Resource Group: grp-Ip-Fs-Msyql
     res-Wmysql3-IP     (ocf::heartbeat:IPaddr2):       Started punnet
     res-Wmysql3-Fs     (ocf::heartbeat:Filesystem):    Started punnet
     res-Mysql  (ocf::heartbeat:mysql): Started punnet
 Master/Slave Set: ms-Drbd-R0
     Masters: [ punnet ]
     Slaves: [ pannier ]
 Clone Set: clo-Ping
     Started: [ punnet pannier ]
```

# Experiences (Heartbeat)

- 0.4.6 (2000) solid, but very limited
- Later versions became less stable, as our usage moved away from what Alan planned
- Not easy to manage (down one node, update software, etc.)
- Good at ensuring at least one node up
- Poor at ensuring at most one node up

# Experiences (corosync)

- The corosync layer and basic pacemaker seems rock solid
- totem token configuration wrong in Debian
- Complexity: using multicast addresses
  - Our problematic networking
  - Once working, it goes on
  - Need to enable encryption for security
- Major improvements over heartbeat

# Experiences (resources)

- 75 agents (at least) in Debian install
- Some duplication
- Varying quality
- Includes Dummy scripts
- Combining and tuning resources is an art
- Many ways to do things, sometimes (apparently) not quite equivalent
- Schwartzkopff's book

# Experiences (STONITH)

- STONITH is not configurable enough
- Where possible, we use application level fencing
- For some reason, we have problems with false positives
- See http://www.ourobengr.com/ha

# Experiences (new RA script)

- Dummy resource very helpful
- Debugging difficult
- For some errors, logging is helpful
- Insert debugging commands into script

# Managment

- CRM
- crm_gui