PostgreSQL on Solaris

PGCon 2007

Josh Berkus, Jim Gates,
Zdenek Kotala, Robert Lor
Sun Microsystems

Agenda

- Sun Cluster
- ZFS
- Zones
- Dtrace
- Service Management Facility (SMF)

Hightly Available HA-PostgreSQL on Sun Cluster

1. Sun Cluster

- Loosely coupled heterogeneous nodes
  - Max. 64 nodes
- Provides single client view of network services or applications
  - databases, web services, file services.
- Highly available and scalable applications
- Capacity for modular growth
- Low entry price compared to traditional hardware fault-tolerant systems.

2. Sun Cluster Continued

- Continuously monitors health of member nodes, networks and storage
- Monitors applications and their dependent system resources, and fails over or restarts applications in case of failures.
- Fault-tolerant hardware systems come at a higher cost because of specialized hardware.

3. Sun Cluster Goals

- Reduce system downtime due to software or hardware failure
- Ensure availability of data and applications, regardless of the kind of failure that would normally take down a single server system
- Increase application throughput by enabling services to scale to additional processors by adding nodes to the cluster
- Perform maintenance without shutting down the entire cluster

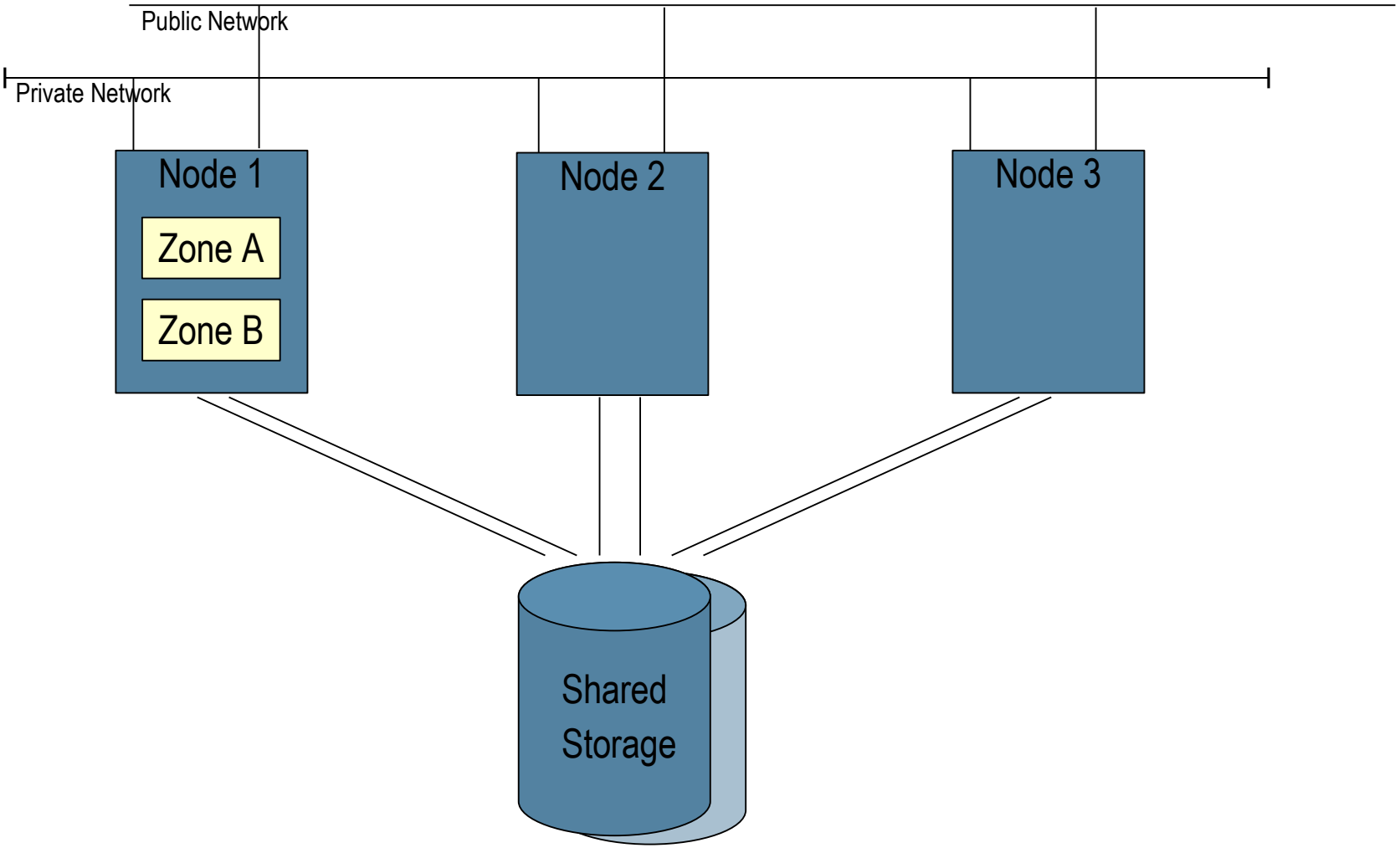
4. Failover and Scalability

- A single cluster can support both failover and scalable applications.
- Failover
  - Cluster automatically relocates an application from a failed primary node or zone to a designated secondary node or zone.
  - Clients might see a brief interruption in service and might need to reconnect after the failover has finished.

5. Failover and Scalability

- Scalability
  - > Leverages the multiple nodes in a cluster to concurrently run an application, thus providing increased performance.
  - > Each node in the cluster can provide data and process client requests.
- HA-PostgreSQL is a failover application

6. Sun Cluster Topology



7. Typical Applications

- HA-Oracle
- Oracle RAC (scalable)
- App Server (HA or scalable)
- Web Server
- HA-NFS
- HA-DNS
- HA-SAP
- HA-Siebel

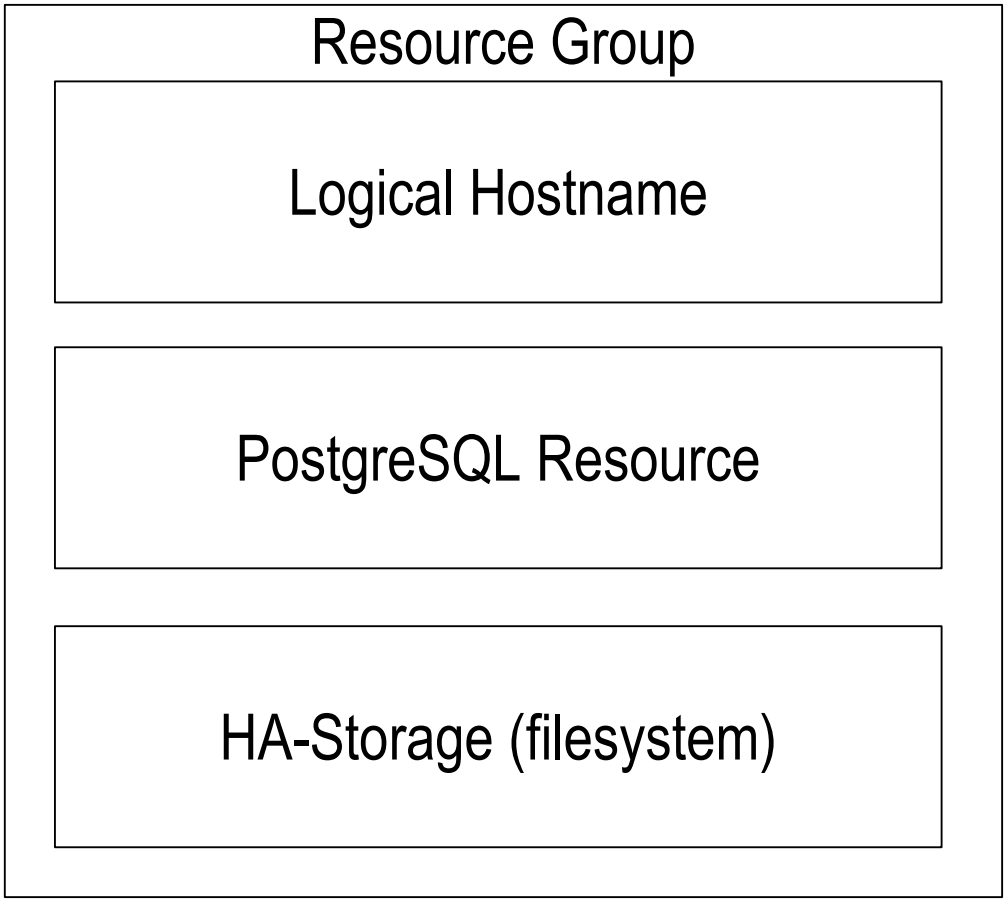
8. Terminology

- Logical hostname
  - IP address, mastered on one node at a time, relocated between nodes
- Resource type
  - Unique name given to a data service object
  - Either failover types or scalable types
  - Associated start, stop, monitor, etc. methods
- Resource
  - An instance of a resource type.
  - Many resources of the same type might exist, each having its own name and set of property values

9. Terminology Continued

- Resource group
  - A collection of resources that are managed as a single unit.
  - All resources must be configured in a resource group.
  - Related and interdependent resources are grouped.
- Data service
  - An application that has been instrumented to run as a highly available resource under control of the Resource Group Manager

10. HA-PostgreSQL Data Service



11. Monitoring

- Starting the resource group
  - Start (plumb up) logical hostname
  - Start (fsck & mount) filesystem
  - Start PostgreSQL (pg_ctl)
  - Monitor PostgreSQL
- Stopping is reverse of start steps
- All HA-PostgreSQL methods implemented as Korne shell scripts

12. Monitoring (continued)

- Monitor method uses ps & psql client tool:
  - Check postmaster running
  - Select datname from pg_database
  - Truncate test table, insert into test table, select from test table
  - Bounded response time (timeout)
- Problems that cause errors initiate:
  - Retry/wait cycle (configurable)
  - Restart of database cluster instance
  - Failover of resource group to another zone or node

13. Monitoring (continued)

- Typical problems that HA-PostgreSQL/Cluster will recover from
  - Public network failure
  - Node crash
  - Database cluster instance crash
  - Accidental database shutdown
  - System memory shortage
  - Backend errors and hangs

14. Monitoring (continued)

- Problems that HA-PostgreSQL can't prevent or recover:
  - Data corruption
  - Multiple failures e.g. All nodes crash, complete loss of storage
  - Database errors not detected by the monitor

15. Restrictions

- Database cluster (PGDATA) must be placed on shared storage
- Server must perform TCP listen on localhost
- Password policy must be *trust* or *password*

16. Availability

- Sun Cluster version 3.2
  - Requires license
  - Not available for earlier versions 3.1 & 3.0
  - Solaris 9 & 10
  - SPARC, x86 and AMD64
- Data service package on SC Agents DVD
  - SUNWscPostgreSQL
- Any PostgreSQL version that includes pg_ctl & psql
- PostgreSQL for Solaris is free
  - 8.1 bundled with Solaris 10 6/06 (u2) onwards
  - 8.2 bundled with Solaris 10 6/07 (u4) onwards

Solaris ZFS –
The last word in filesystems

ZFS Features

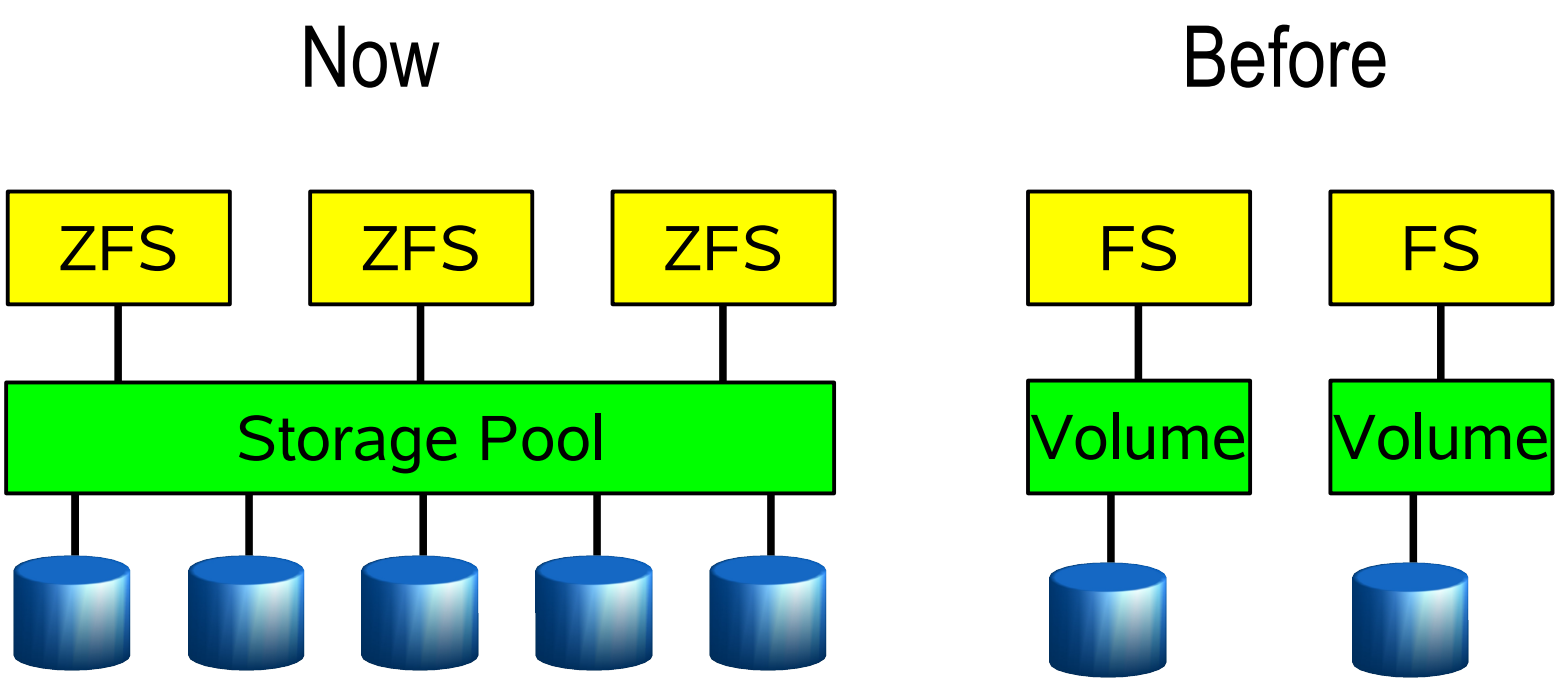
- Pooled Storage Model
- Always consistent on disk
- Protection from data corruption
- Live data scrubbing
- Instantaneous snapshots and clones
- Fast native backup and restore
- Highly scalable
- Built in compression
- Simplified administration model

Design Principles

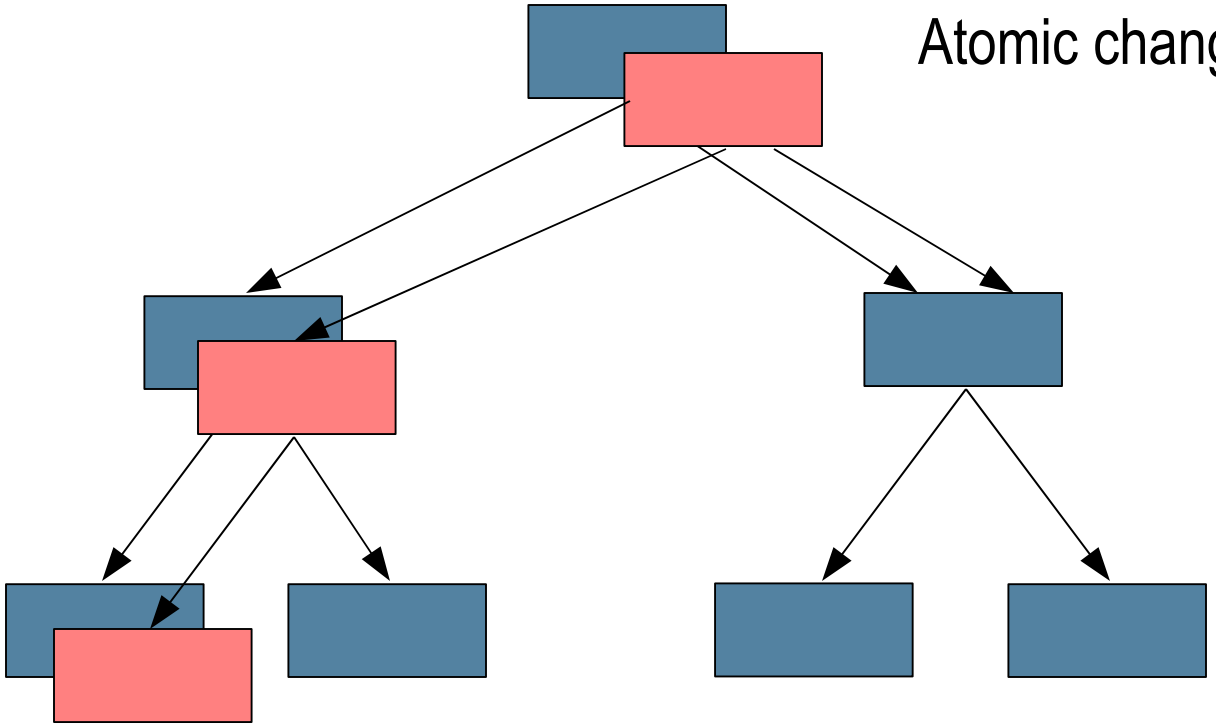
- Completely new design
  - 128-bit
  - Copy on write (COW)
- Pooled storage
  - ZPOOL/ZFS does for storage what VM does for memory
- End-to-end data integrity
  - Validates the entire I/O path
- Transactional operation
  - Keep things always consistent on disk
  - Removes almost all constraints on I/O order

ZFS

- The World's First 128-bit File System
  - Zetabyte = 70 bit (a billion TB)
  - ZFS capacity = 256 quadrillion ZB
- Pooled Storage



Copy On Write



Atomic change of root node

ZFS basic

- Create pool
  - # zpool create pgsql c1t2d0
- Create filesystem
  - # zfs create -b 8192 pgsql/data
  - # zfs set mountpoint=/var/postgres/data pgsql/data
- And more ...
  - # zfs create pgsql/data/base
  - # zfs create pgsql/data/tablespace_1
  - # zfs create pgsql/data/tablespace_2

ZFS basic II.

- Compress data
  - # zfs set compression=on pgsql/data
- Set quota
  - # zfs set quota=10g pgsql/data/tablespace_1
- Reserve space
  - # zfs set reservation=10g pgsql/data/tablespace_1
- No space?
  - # zpool add pgsql c5t1d0 c6t1d0
- NFS-export
  - # zfs set sharenfs=rw pgsql/data

ZFS Advance

- Snapshot
  - Read only point-in-time copy of filesystem
  - Access through .zfs/snapshot in root of each filesystem
  - # zfs snapshot pgsql/data@tuesday
  - # zfs rollback pgsql/data@tuesday
- Clone
  - Writable copy of snapshot
  - # zfs clone pgsql/data@tuesday pgsql/data_test

ZFS Advance II.

- Backup/Restore
  - Based on snapshot
  - Full or incremental
    - #zfs send psql/data@tuesday > /var/postgres/backup/back.full
  - Incremental
    - #zfs send -i psql/data@tuesday psql/data@wednesday> /var/postgres/backup/back.inc
  - Possible to do remote replication
    - # zfs send -i psql/data@11:31 psql/data@11:32 | ssh host zfs receive -d /var/postgres/data

Solaris Containers

Containers: Zones + Resource Control

- Run multiple applications on one system
  - Improve utilization
  - Reduce management overhead
- Isolate applications from:
  - Faults
  - Intrusion
- Resource control
  - CPU
  - Memory
  - Processes

Containers Demo

Solaris Dynamic Tracing

Introducing DTrace

- Allows for dynamic instrumentation of the OS and applications
- Available on stock systems - typical system has more than 30,000 probes
- Dynamically interpreted language allows for arbitrary actions and predicates

Introducing Dtrace, cont.

- Designed explicitly for use on production systems
- Zero performance impact when not in use
- Completely safe - no way to cause panics, crashes, data corruption
- Powerful data management primitives eliminate need for most postprocessing
- Unwanted data is pruned as close to the source as possible

Dtrace Demo

Solaris Service Management Facility

SMF

- Defines relationships among:
  - Applications
  - Solaris components
- Simplifies administration
  - Consolidates “application profile”
  - Helps manage components
- Increases service reliability
  - Detects service outages quickly
  - Recovers services accurately

SMF Demo

Further Information

- OpenSolaris Communities
  - http://www.opensolaris.org
- Sun Cluster home page
  - http://www.sun.com/software/solaris/cluster/index.xml
- HA-PostgreSQL page
  - http://www.sun.com/download/products.xml?id=44774401
- PostgreSQL on Solaris
  - http://www.sun.com/software/solaris/postgresql.jsp
- PostgreSQL on Solaris How To Guide
  - http://www.sun.com/software/solaris/howtoguides/postgresqlhowto.jsp
- Sun Support Service Plans
  - http://www.sun.com/service/osdb/index.xml

Josh Berkus josh.berkus@sun.com
Jim Gates jim.gates@sun.com
Zdenek Kotala zdenek.kotala@sun.com
Robert Lor robert.lor@sun.com