Distributed Peer-to-Peer Control in Harness

C. Engelmann, S.L.Scott, G.A.Geist

Oak Ridge National Laboratory, USA

http://www.csm.ornl.gov/harness/

{engelmannc|scottsl|gst}@ornl.gov

What is Harness?

- Successor of PVM (parallel virtual machine).
- Conceived as DVM (distributed virtual machine).
- Provides high-availability.
- Supports plug-in mechanism.
- Enables collaborative computing.
- Collaborative effort between:
  - Oak Ridge National Laboratory, USA
  - University of Tennessee, USA
  - Emory University, USA

What is distributed control?

- Global state control in a distributed system with failures:
  - Every machine is able to change the global state.
  - Global state replication to provide fault-tolerance.
  - Global state change verification to provide consistency.

- Management of a distributed state database:
  - Every machine has a complete copy of the global state.
  - Global state changes are transactions, which are ordered, executed, and committed or rejected.

Distributed Control in Harness?

- Controls global DVM state:
  - Member configuration and DVM membership.
  - Plug-in loading, unloading and configuration.
- Provides high-availability:
  - DVM survives until at least one member is alive.
  - Hot-standby or warm-standby plug-ins.
- Supports event distribution:
  - Member or plug-in failure notification.
  - Member or plug-in state change notification.

Distributed Peer-to-Peer Control

- Scalable peer-to-peer ring.
- All members with the same global state form one ring.
- Messages are forwarded only in one direction.
- Transactions are ordered, executed and committed or rejected using group communication.
- Connections are persistent.
- TCP/IP provides fault detection and ensures message order on the ring.

Group Communication

Reliable Broadcast:

State changes are broadcasted reliably.

- Messages go twice the ring (2-phase commit).
- The last phase 2 and all phase 1 messages are sent again by a member to recover from faults.
- Doubled messages are filtered by the receiver using a hop counter contained by every message.

Group Communication

Atomic Broadcast:

Reliably broadcasted state changes are globally ordered.

- Message numbering without timestamps.
- Message sorting without blocking.
- No starvation or denial of service due to fair share.

Group Communication

Distributed Agreement:

All members agree on a state change.

- Collective communication combines state change execution results from all members to a final result.
- Final execution results are broadcasted reliably.
- Messages go 3 times around the ring
(2 interleaved 2-phase commits: collection & final result).

Group Communication

Membership:

All members agree on an initial state.

- Every new member receives the current global state.

All members have a linear history of state changes.

- Atomic Broadcast of state changes (two phases).
- Distributed Agreement on execution results (three phases).
- State change commit depending on final result.

Conclusions

- Distributed peer-to-peer control:
  - Fault-tolerant distributed global state control.
  - Scalable group communication (2n-5n).
  - Avoidance of starvation and denial of service.
- Advantages for Harness:
  - Scalable global state control and event notification.
  - High-available distributed virtual machine.
  - Distributed plug-in management.

Distributed Peer-to-Peer Control in Harness

C. Engelmann, S.L.Scott, G.A.Geist

Oak Ridge National Laboratory, USA

http://www.csm.ornl.gov/harness/

{engelmannc|scottsl|gst}@ornl.gov