CSE 791: High-Performance Distributed Computing

T/Th, 4:00-5:20

Link 105

Fall, 1999

Introduction to HPDC and Clustering

High-Performance

- Time matters
- It is important to:
  - have a high-performance substrate
    - fast machines, fast disks, fast networks
    - latency can be as least as important as bandwidth
  - deliver that performance to the application
- We will focus on the latter

What are (Heterogeneous) Distributed Systems?

- Abstract definition:
  - machines with differences in architecture, configuration, and implementation
  - no shared memory (message passing)
- Operational definition:
  - clusters of workstations/PCs
  - wide-area collections of autonomous domains

What is a Cluster?

- We don't even have agreement within the cluster computing community.
- Cf. http://www.dgs.monash.edu/~rajkumar/tfcc
- Various definitions, none of which exactly encompass a cluster

Pfister's Definition

- "A cluster is a type of parallel or distributed system that consists of a collection of interconnected whole computers used as a single, unified computing resource."
- "Whole computer is meant to indicate a normal, whole computer system that can be used on its own…"

But...

- What about diskless workstations, or those with minimalist OS?
- Why does the entire cluster have to be used as a single resource?
- At least it's a short definition…
- See the TFCC-L archives for more definitions.

For Our Purposes

- Pfister's definition is close enough, modulo my objections
- We'll use the supreme court pornography test: we'll know a cluster when we see it.

Some Problems in Heterogeneous Distributed Systems

- Resource management: getting jobs to where they'll run best.
- Performance: Delivering the underlying capability to the application.
- Security: providing assurance to resource owners that their access policies will be respected

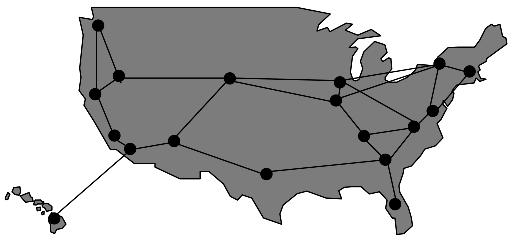
Distributed Computing Research Projects

- Orange Grove & Centurion: cluster-based distributed hardware
- HOSS: OS-layer for heterogeneous clusters
- IBID: Interface-Based Intrusion Detection
- WARMStones: benchmarking scheduling algorithms
- Wide-Area Network Management

Distributed Cluster Computing

Distributed Cluster Computing

- Orange Grove (SU cluster)
- Centurion (UVa cluster)
- Connection via vBNS



The Orange Grove

- Cluster-based supercomputer
- 16 DEC Alpha workstations
- 48 Dual-Processor Intel machines
- Myrinet, Fast Ethernet
- Linux

Centurion

- 128 Alphas
- 128 Pentium-based dual processors
- Fast Ethernet/Myrinet
- Linux
- Paid for by NSF (Co-PI) and ONR DURIP (PI)

HOSS

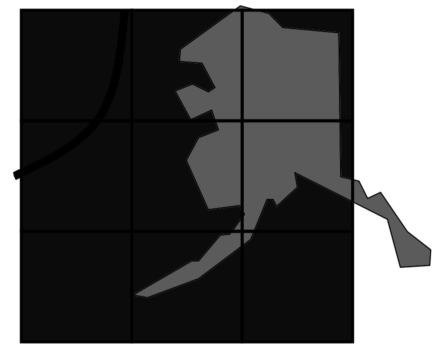
HOSS

- Focus on scalable distributed operating systems for clusters
- Based on Linux (Intel & Alpha)
- Flexible support for delivering the performance of underlying hardware to the application
- Example: threads

Threads in a Clustered System

- MPI programs use spatial decomposition
- Threaded programs decompose on control (requires shared memory)
- What about mixing the two?
- Example: weather modeling

Weather Modeling: Front Movement



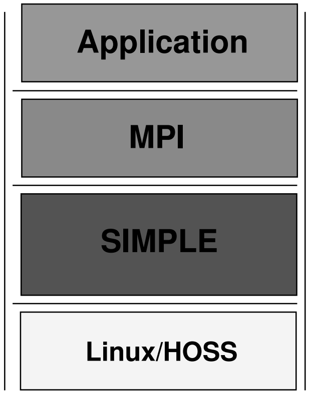
Weather Modeling

Weather Modeling

Weather Modeling

Thread Layering

- MPI: mixed thread model, Mike Heroux, Sandia
- SIMPLE: run-time system, David Bader, UNM ECE
- HOSS: dynamic thread support, Steve Chapin



Intrusion Detection

Intrusion Detection

- Intrusion detection determines when an outsider is accessing the system in a harmful way
- Defense in depth required: no one wall will suffice
- Devise complementary modules which together provide security

Current Forms of Intrusion Detection

- Host-based
  - ◆host monitors application actions, files, etc.
  - ◆coordinates with central monitor
- Router-based
  - ◆monitors network traffic
  - ◆can detect some problems

Problems with Coverage

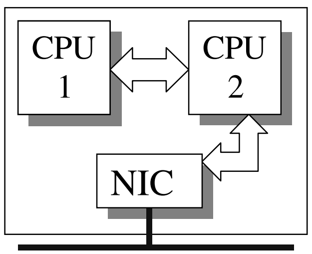
- Router-based schemes
  - ◆bottleneck
  - ◆need information from hosts
  - ◆doesn't protect from local attack
  - ◆may require others to play well
- Host-based schemes
  - ◆must still deal with network issues

A New Module: Interface-Based Intrusion Detection

- Put more smarts in the NIC
- Can ease load on both routers and hosts
  - ◆routers don't have to do as much processing on each packet
  - ◆fewer network issues for hosts
- Completely local defense model

Planned Approach

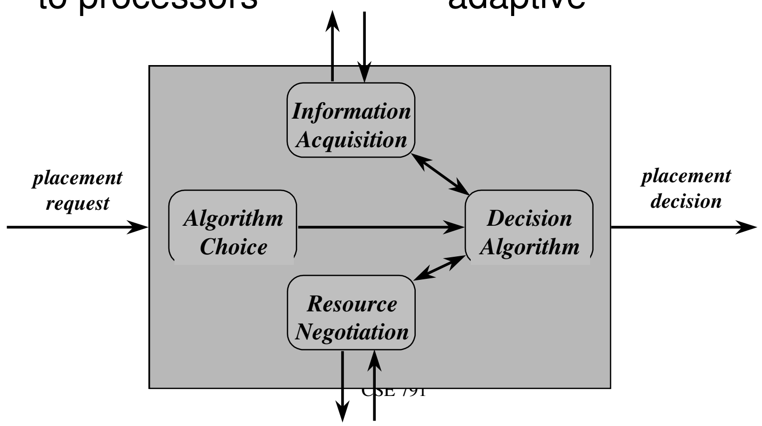
- Use dual-processor Pentium boxes running Linux
- One processor acts as normal
- Second processor handles all network I/O, acts as smart extension to NIC



WARMStones

Global Scheduling

- mapping of task sets to processors
- static, dynamic, adaptive



The Comparison Problem

- Hundreds of scheduling algorithms
- Primarily theoretical results
- No practical basis for comparison

We need some way of judging the quality of scheduling work

The Solution: Benchmarks

- The “standard” answer (Whetstones, Dhrystones, Hartstones)
- Basis for comparing scheduling algorithms and systems
- Standard task suites covering the interesting problem space characterize the performance of scheduling algorithms

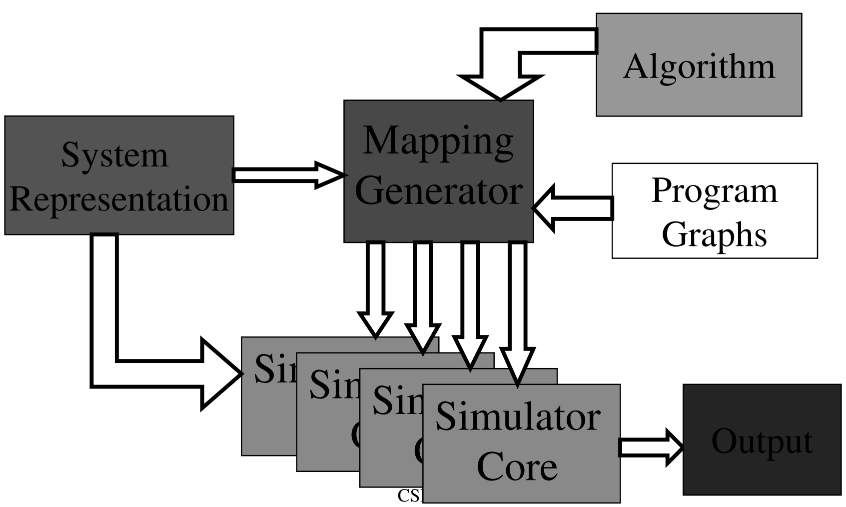
The Proposal: WARMstones

- Draw from the best of several current simulators/systems, e.g.:
  - ◆K9: parallel machines
  - ◆Osculant: distributed systems
  - ◆GR: graph representation
  - ◆PP-MESS-SIM: message passing
  - ◆MESSIAHS: implementation toolkits

WARMstones II

- Standard program graph representation
- Multiprocessor/multicomputer specification
- Discrete event simulator
- Local and global scheduling policy toolkits (with defaults)
- Simulate scheduling algorithms on multiple task suite/machine pairs

WARMstones Architecture



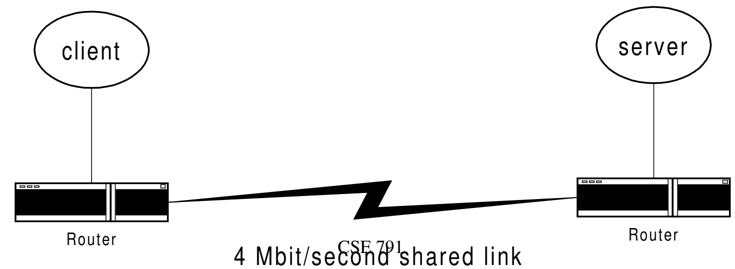
Wide-Area Network Management

Wide-Area Network Management

- Remember our distributed clusters
- We have a good handle on the intra-cluster networking
- But we can't yet allocate resources between clusters
- Luckily, new switches are on the way

A Simple Example

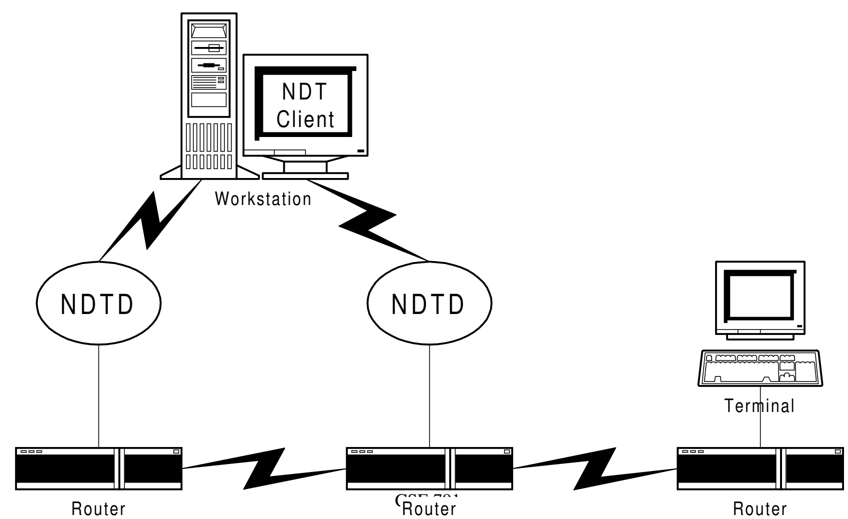
- Want to manage a single, shared link
- "Priority" user wants 2 Mbps
- Rest of traffic gets the other channel
- cf. ATM



Hardware Requirements

- We can build software to allocate resources wrt QoS
- But we need hardware support for policing
  - ◆RSVP, DiffServ, MPLS, etc.
- New routers have such support
  - ◆Avici, Juniper, Cisco, Lucent, etc.

More General Usage



Use of WANM

- Our modules control local resources
- Form building blocks for higher-level metacomputing/grid software
  - Globus
  - Legion
  - etc.

Summary

- We have many interesting problems to solve in distributed, heterogeneous systems
- We have cool toys
- We will be getting more funding to support Research Assistants

Latency vs. Bandwidth

- Ryder truck of DAT tapes
  - 2G per tape, 3 cubic inches per tape
  - Ryder truck holds
- Latency:
  - C = 186k mps
  - 3k miles SF to NY
  - 186 / 3 => 1/30 second round trip (33 ms)