Cray SV2 Scheduling & Placement

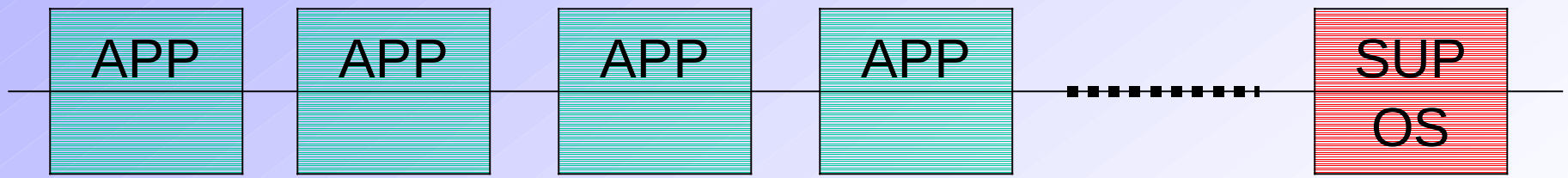
- **Stephan Gipp**
- skg@cray.com

Overview

- Resource Flavor Concept
- Application
- Psched
- Application Launch

Resource Flavor Concept

User Level View of System:
Application Nodes and Support Node(s)



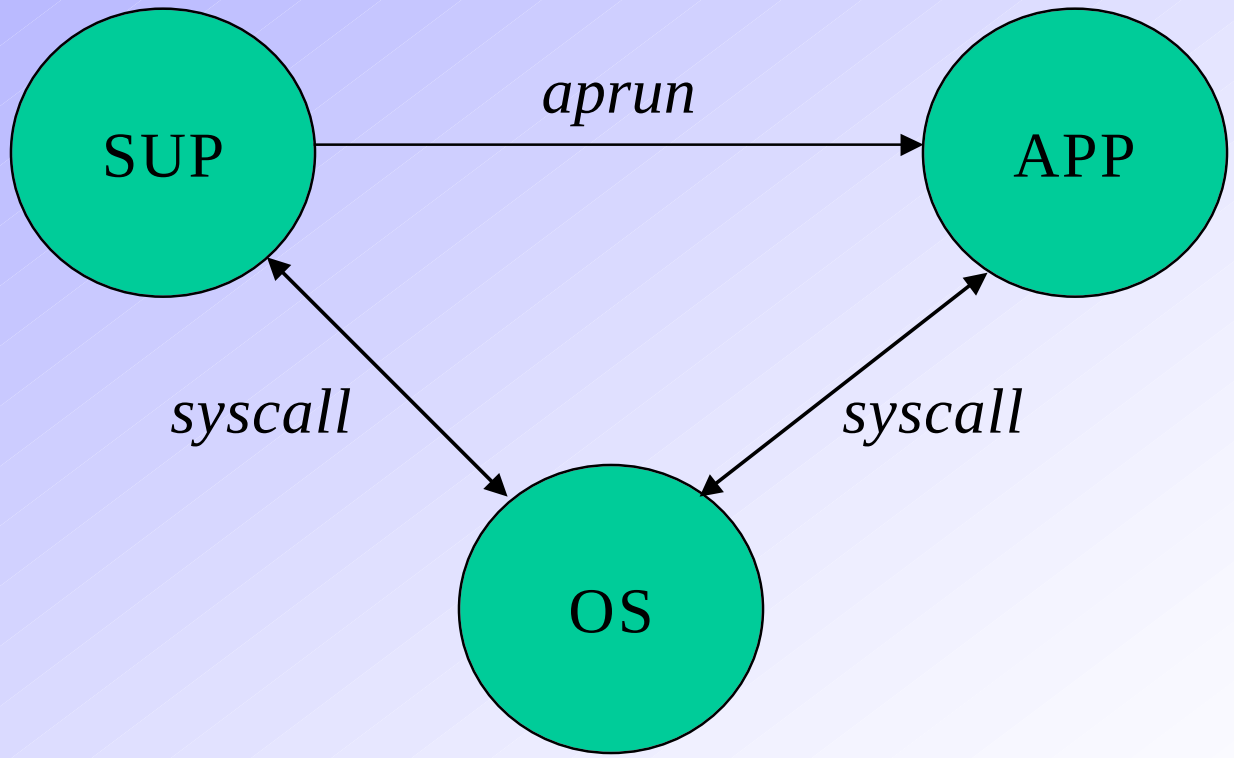
Kernel View of System:
Unflavored Nodes and OS Node(s)

Resource Flavor Concept

- Flavors
  - OS, SUPPORT, APPLICATION
- Resources
  - Processors and Memory
- Providers
  - Nodes
- Consumers
  - Processes and Threads

Resource Flavor Concept

Process and Thread Flavor Transitions



Application

- **Application**
  - User Defined Processes
  - Space Share Scheduling
  - Placed Memory Management
- **Support**
  - Regular UNIX Processes
  - Time Share Scheduling
  - Virtual Memory Management

Application

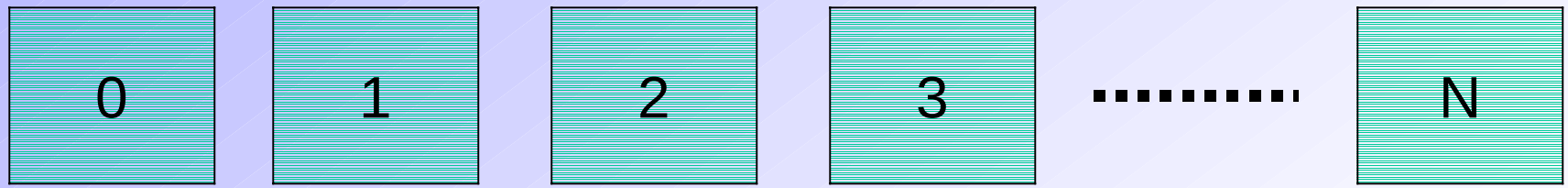
- Group of UNIX Processes
- Identical Binary
- New Memory Sharing Ability
- Common Control (`apid`)
- Accelerated or Flexible Placement Mode

Accelerated Placement Mode

- Remote Translation Table (RTT)
  - + Scalable Address Translation
  - + Consistent Performance
  - - Cray T3E Style Placement

Accelerated Placement Mode

Effective System Topology



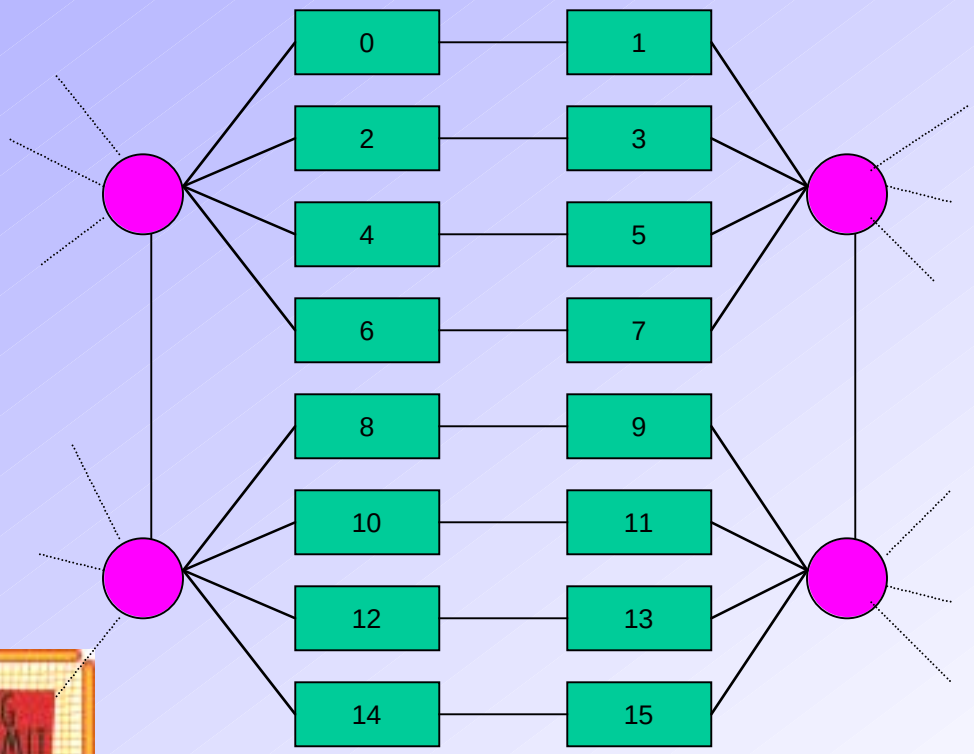
CRAY

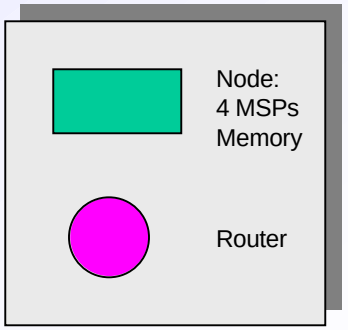
Flexible Placement Mode

- Translation Lookaside Buffer (TLB)
  + Flexible Placement
  - Variable Performance

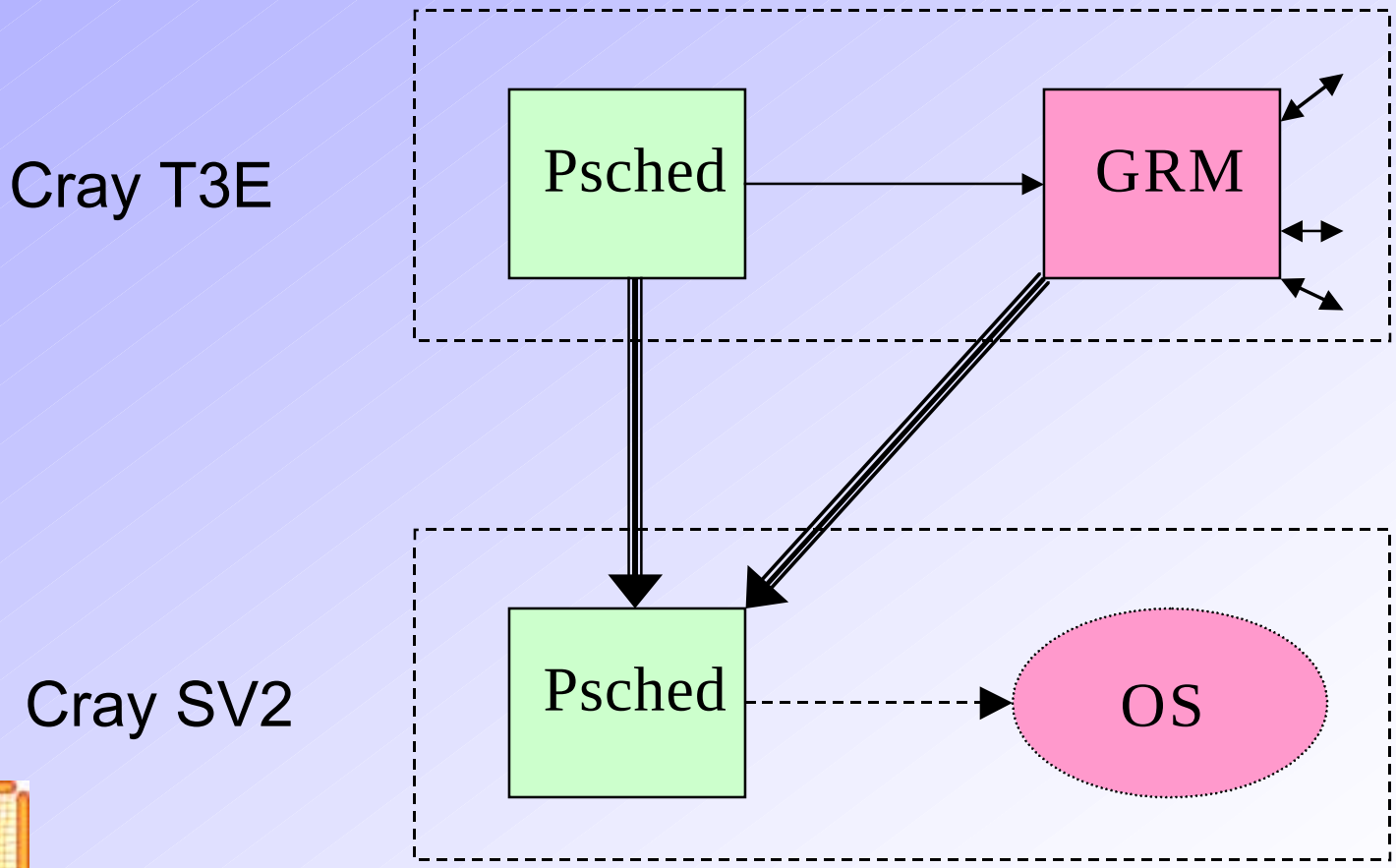
Flexible Placement Mode

Effective System Topology





Psched



Psched Concepts

- Region
  - Describes Nodes
- Domain
  - Describes Scheduling Partitions
- Gates
  - Static Access Control
- Limits
  - Dynamic Access Control

Psched Gates and Limits

- **Attributes**

  Prime, Interactive, Batch, ACX, FLX, MSP, Single, Width, Memory, Time, Hard Label, Soft Label, User ID, Group ID, Account ID

- **Oversubscription (Limits)**

  –Memory, Parties

CRAY

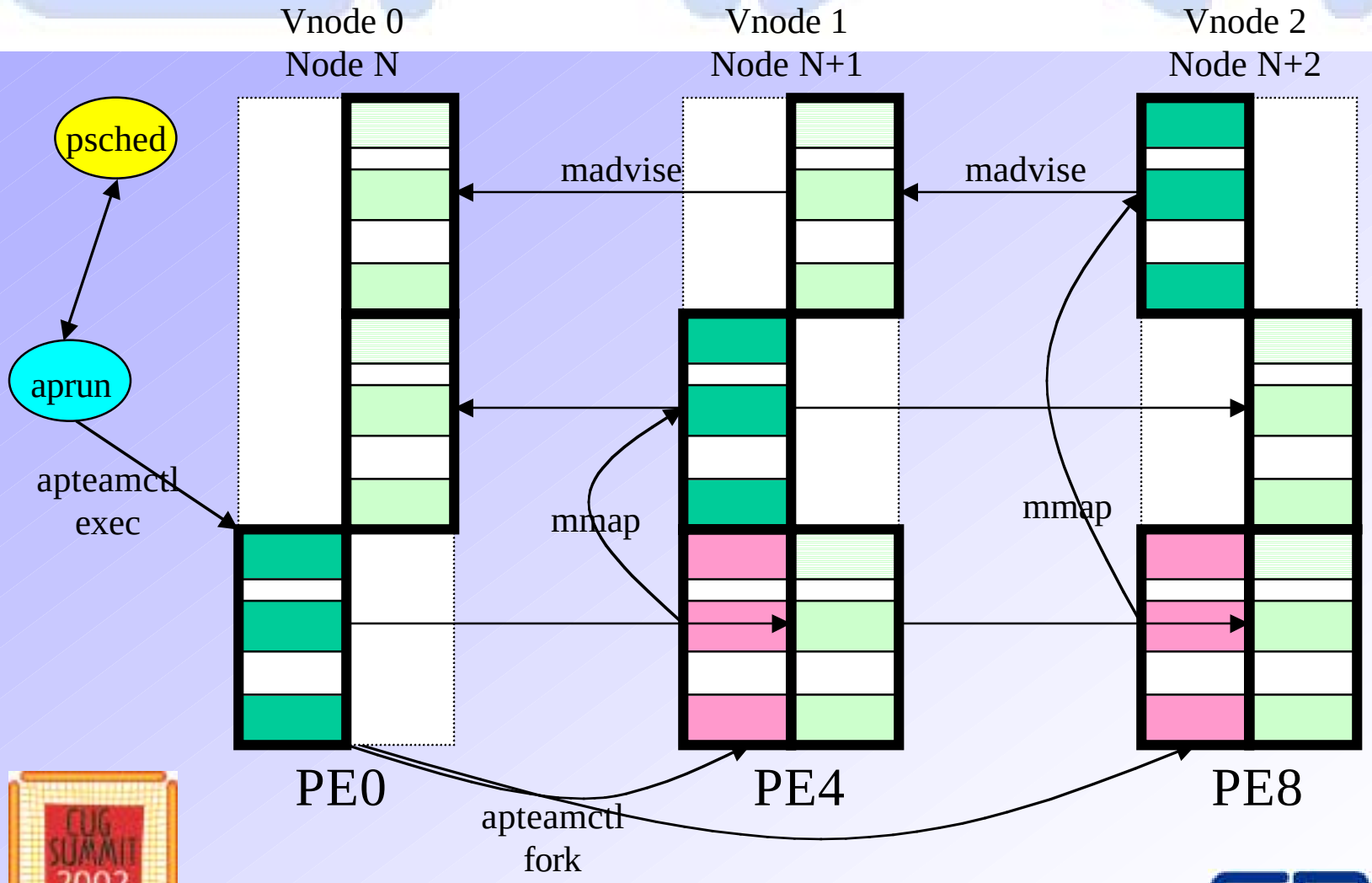
Psched Scheduling

- **Load Balancer & Gang Scheduler**
  - **Like Cray T3E**
    - **Dynamic Load Balancing, Balancing Rules, Prime Applications, Gang Scheduling, Oversubscription Control**
  - **Initial Placement (Load Balancer)**
  - **Launch Time (Gang Scheduler)**

Application Launch

- aprun / mpirun
- Psched:
  - Sets Placement Parameters
  - Sets Launch Time
- libc Startup code:
  - Launch Siblings
  - Initialize Memory Layout

Application Launch



Summary

- Cray T3E Style Scheduler on a NUMA Architecture
- Repeatable Performance for Applications
- Maximum Performance for Applications