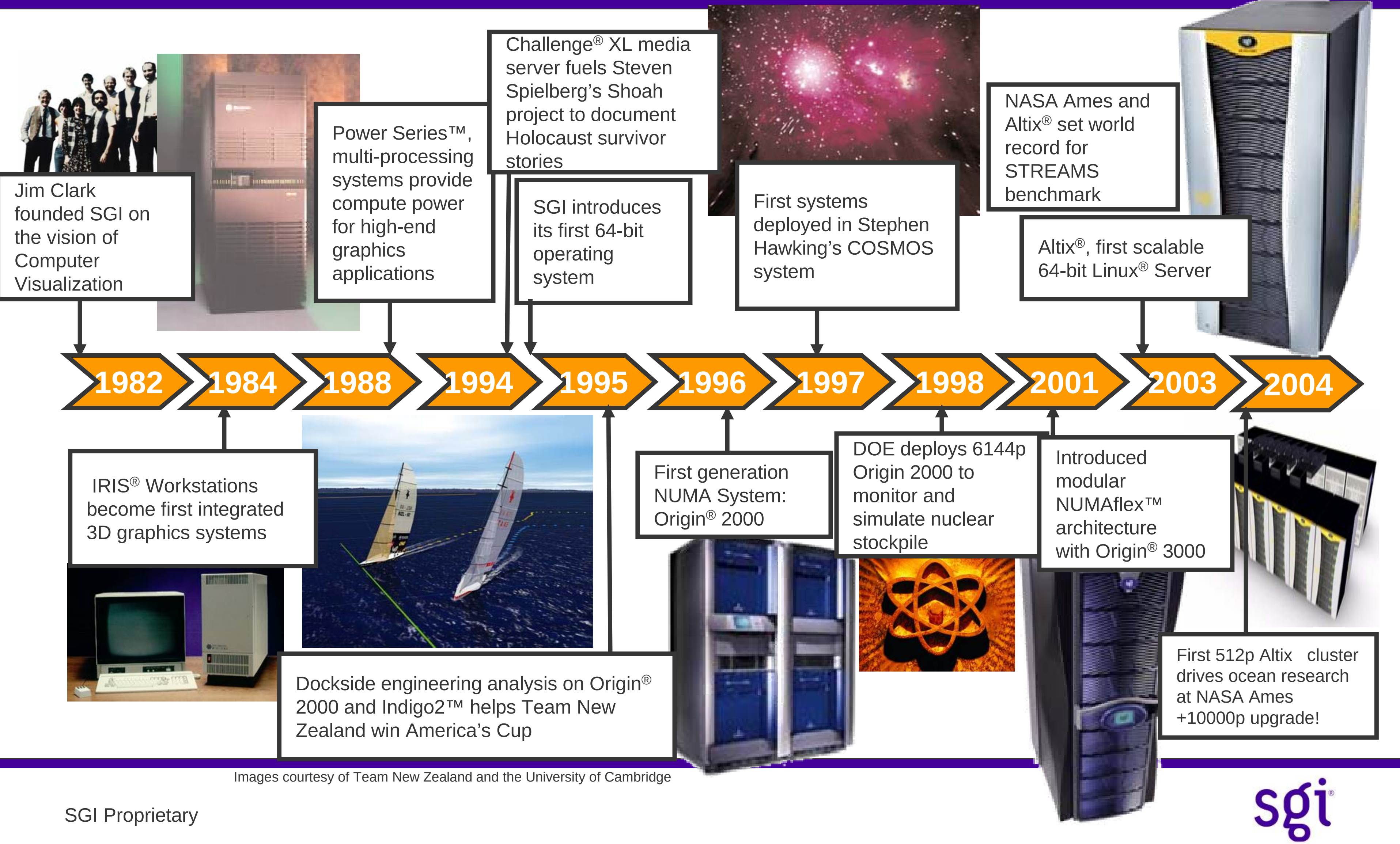
SGI Multi-Paradigm Architecture

Michael Woodacre

Chief Engineer, Server Platform Group

woodacre@sgi.com

A History of Innovation in HPC



Jim Clark founded SGI on the vision of Computer Visualization

Power Series™, multi-processing systems provide compute power for high-end graphics applications

Challenge® XL media server fuels Steven Spielberg's Shoah project to document Holocaust survivor stories

SGI introduces its first 64-bit operating system

First systems deployed in Stephen Hawking's COSMOS system

NASA Ames and Altix® set world record for STREAMS benchmark

Altix®, first scalable 64-bit Linux® Server

1982 1984 1988 1994 1995 1996 1997 1998 2001 2003 2004

IRIS® Workstations become first integrated 3D graphics systems

Dockside engineering analysis on Origin® 2000 and Indigo2™ helps Team New Zealand win America's Cup

First generation NUMA System: Origin® 2000

DOE deploys 6144p Origin 2000 to monitor and simulate nuclear stockpile

Introduced modular NUMAflex™ architecture with Origin® 3000

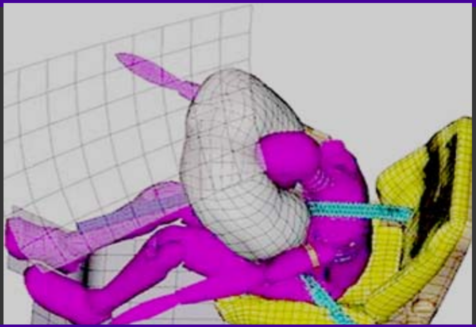
First 512p Altix cluster drives ocean research at NASA Ames +10000p upgrade!

Images courtesy of Team New Zealand and the University of Cambridge

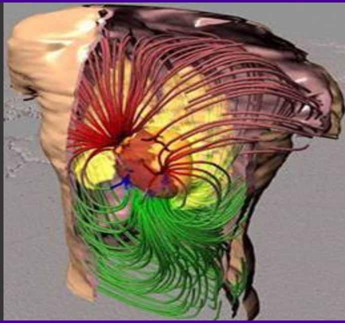
Over Time, Problems Get More Complex, Data Sets Exploding

Bumper, hood, engine, wheels

Entire car



E-crash dummy

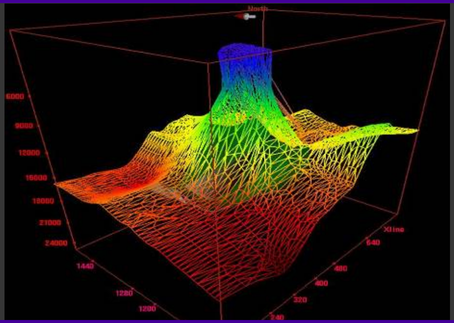


Organ damage

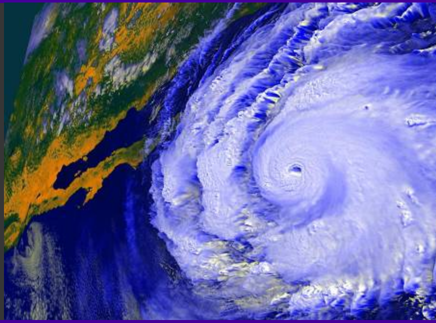
This Trend Continues Across SGI's Markets

Improve design
& manufacturing

Improve patient safety



Improve oil exploration

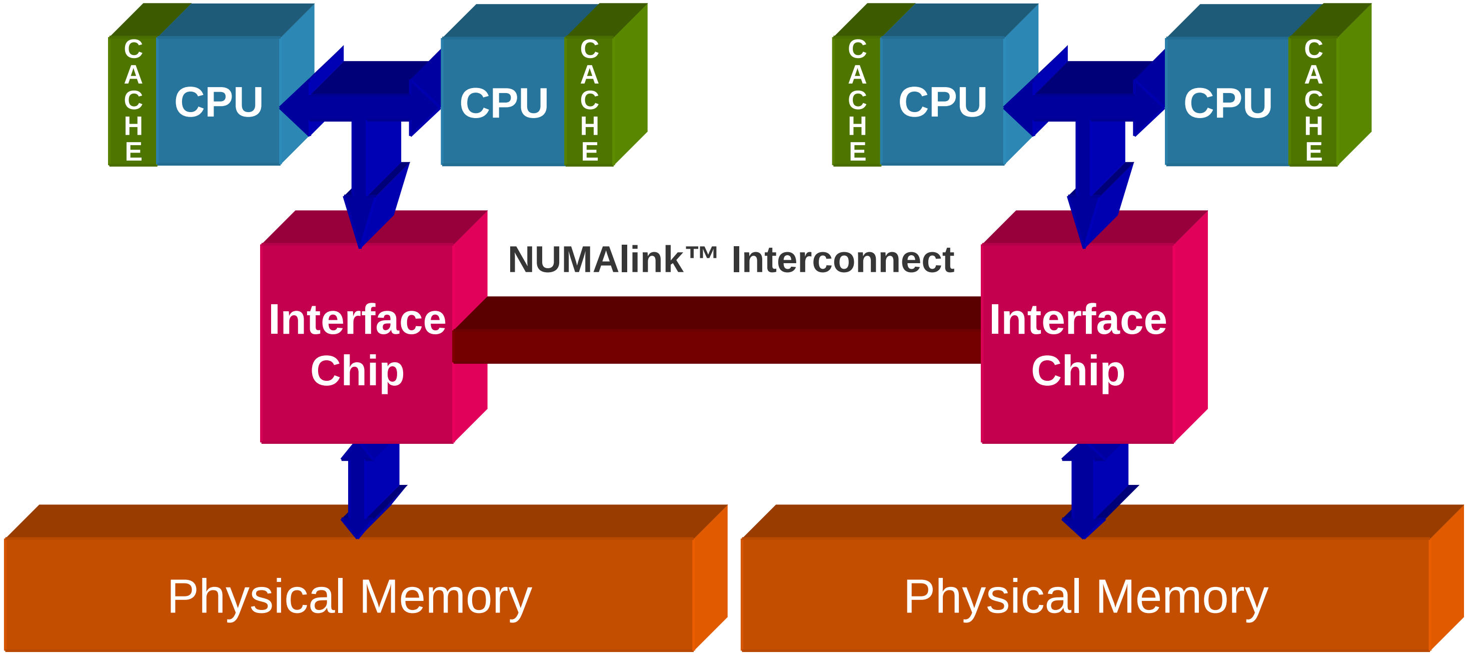


Improve hurricane prediction

First Row Images: EAI, Lana Rushing, Engineering Animation, Inc, Volvo Car Corporation, Images courtesy of the SCI, Second Row Images: The MacNeal-Schwendler Corp , Manchester Visualization Center and University Department of Surgery, Paradigm Geophysical, the Laboratory for Atmospheres, NASA Goddard Space Flight Center.

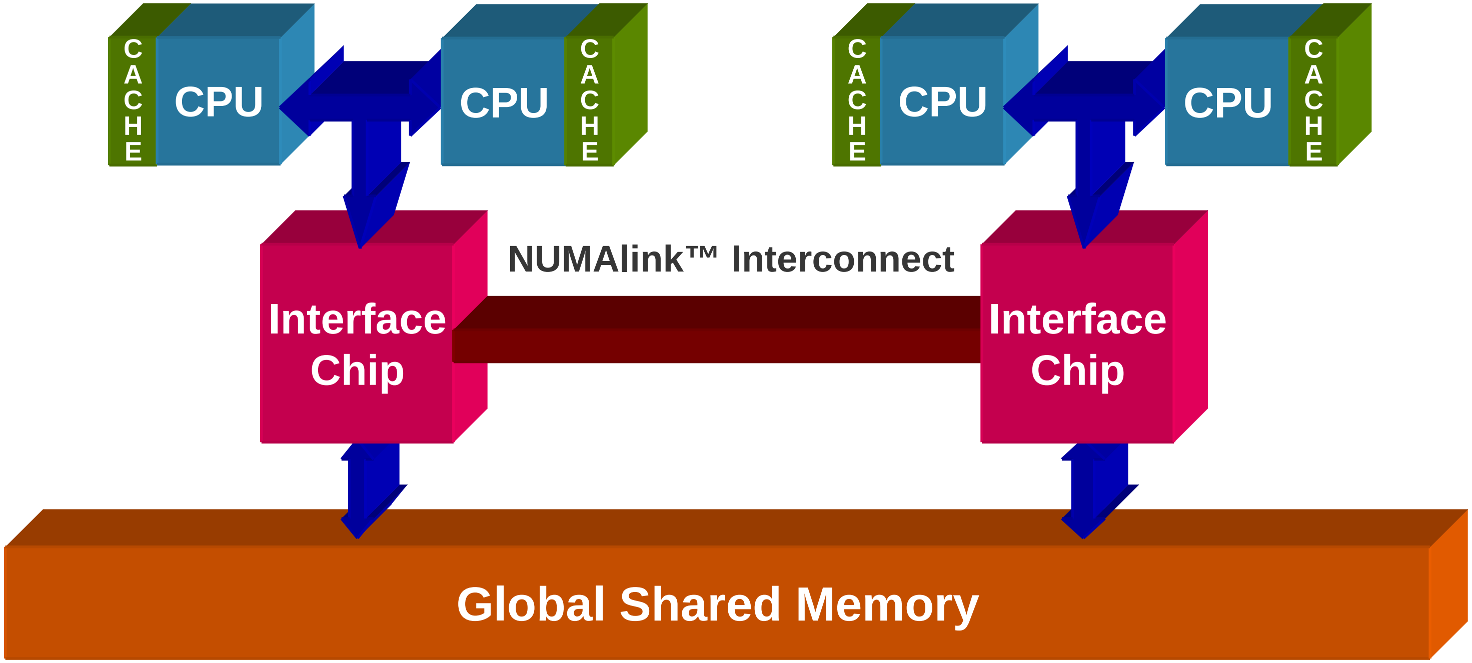
SGI Scalable ccNUMA Architecture

Basic Node Structure and Interconnect



SGI Scalable ccNUMA Architecture

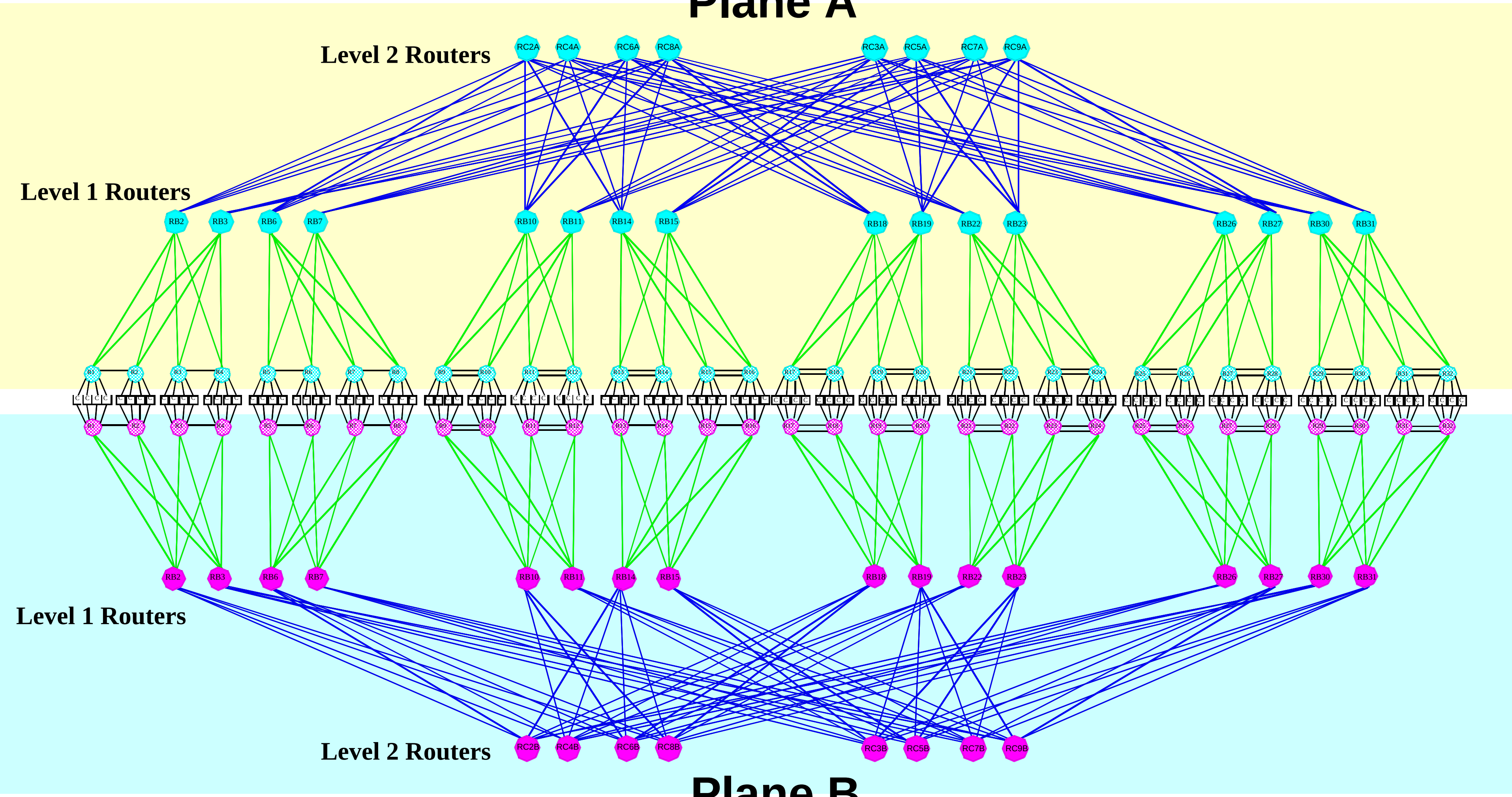
Basic Node Structure and Interconnect



Logical Layout - 8TB

Altix 128 Processor 8TB - 1.6GB/s Uniform Memory Bandwidth

Plane A

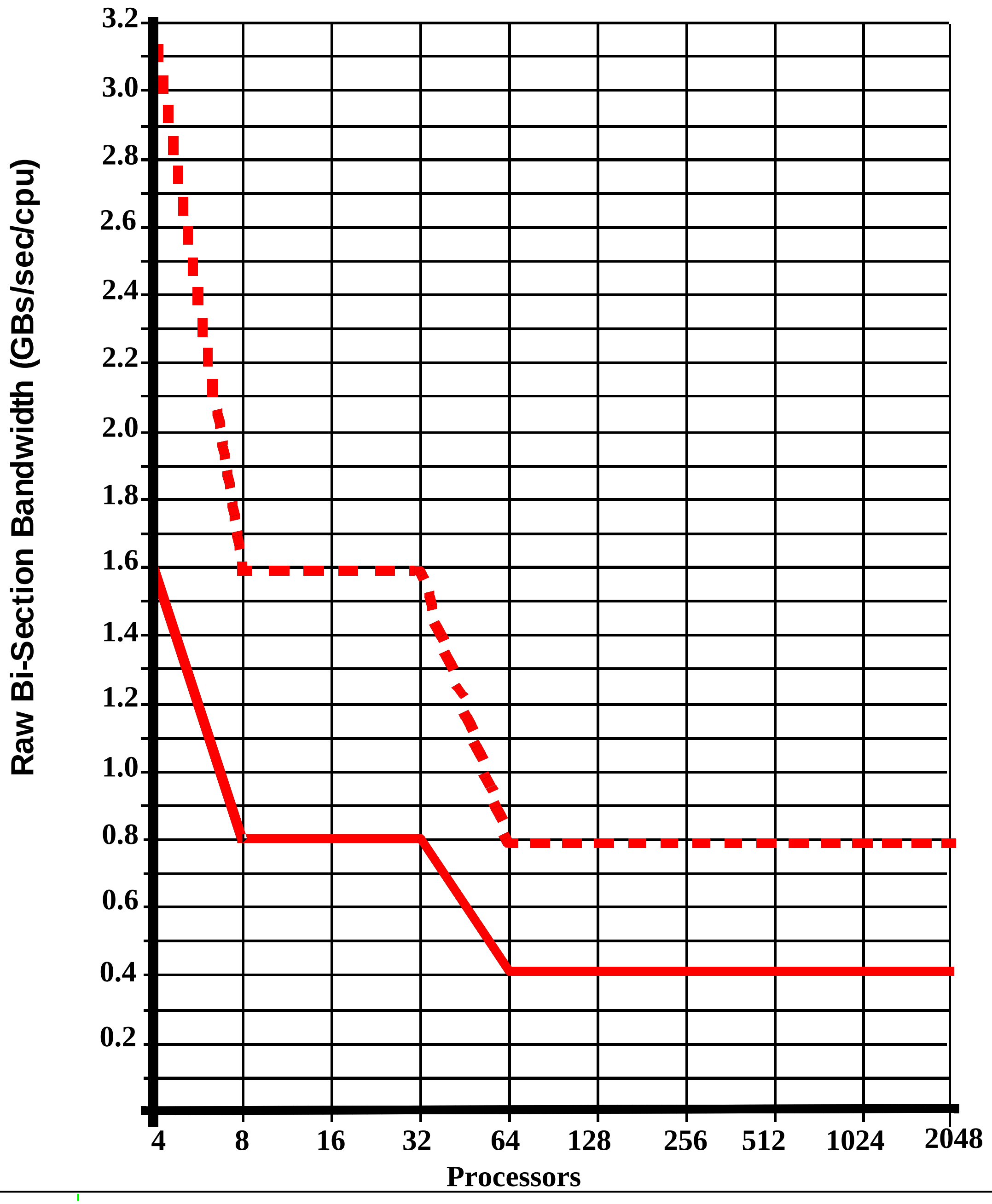


Plane B

Interconnect Topology

Bi-Section Bandwidth Profiles

GBs/sec/cpu



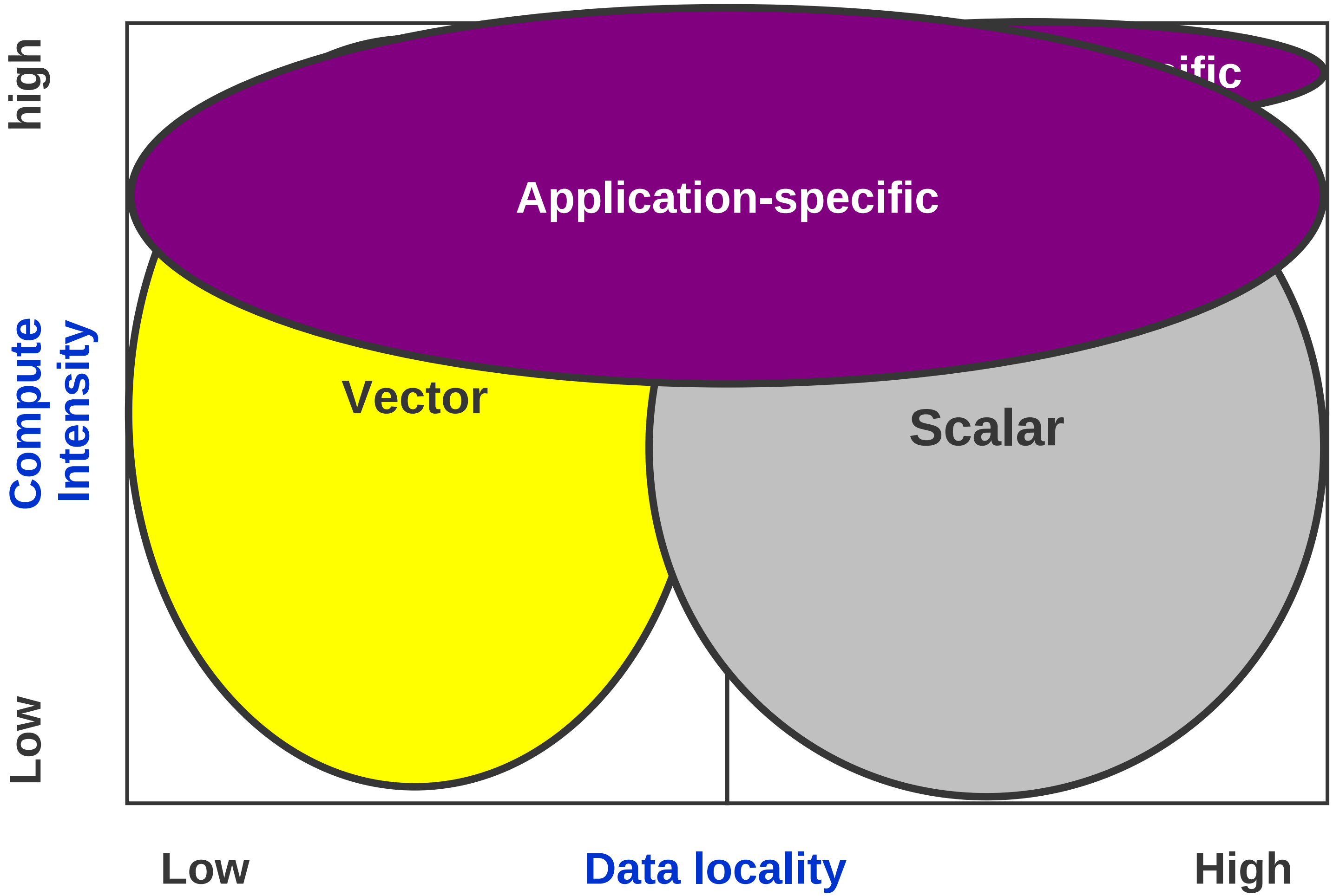
▬ Dual Plane - NL3 router - 8 port router bricks

▪ ▪ ▪ ▪ Dual Plane - NL4 router - 8 port router bricks

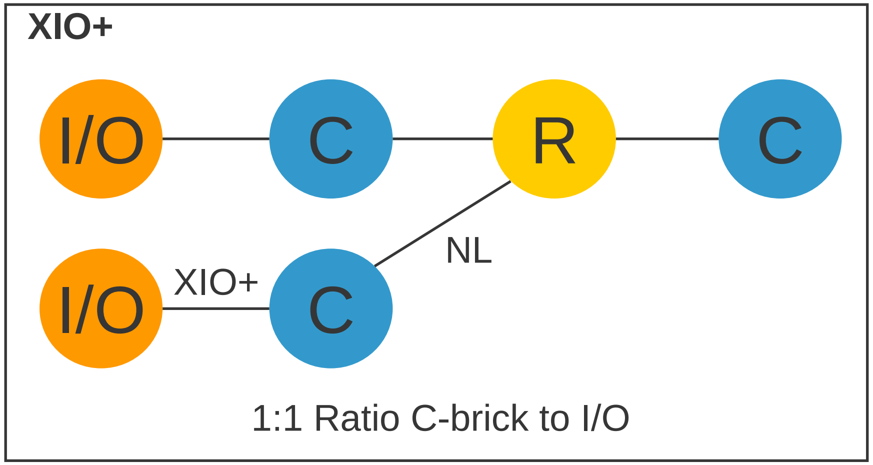
Examples of Single-Paradigm Architectures

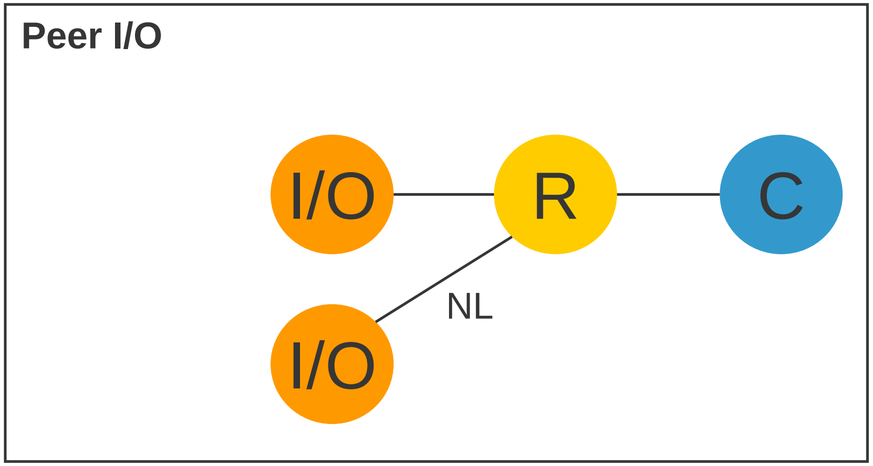
Scalar	**Vector**	**App-Specific**
Intel Itanium	Cray XI	Graphics - GPU
SGI MIPS	NEC SX	Signals - DSP
IBM Power		Prog'ble - FPGA
Sun SPARC		Other ASICs
HP PA		

Paradigms to Applications



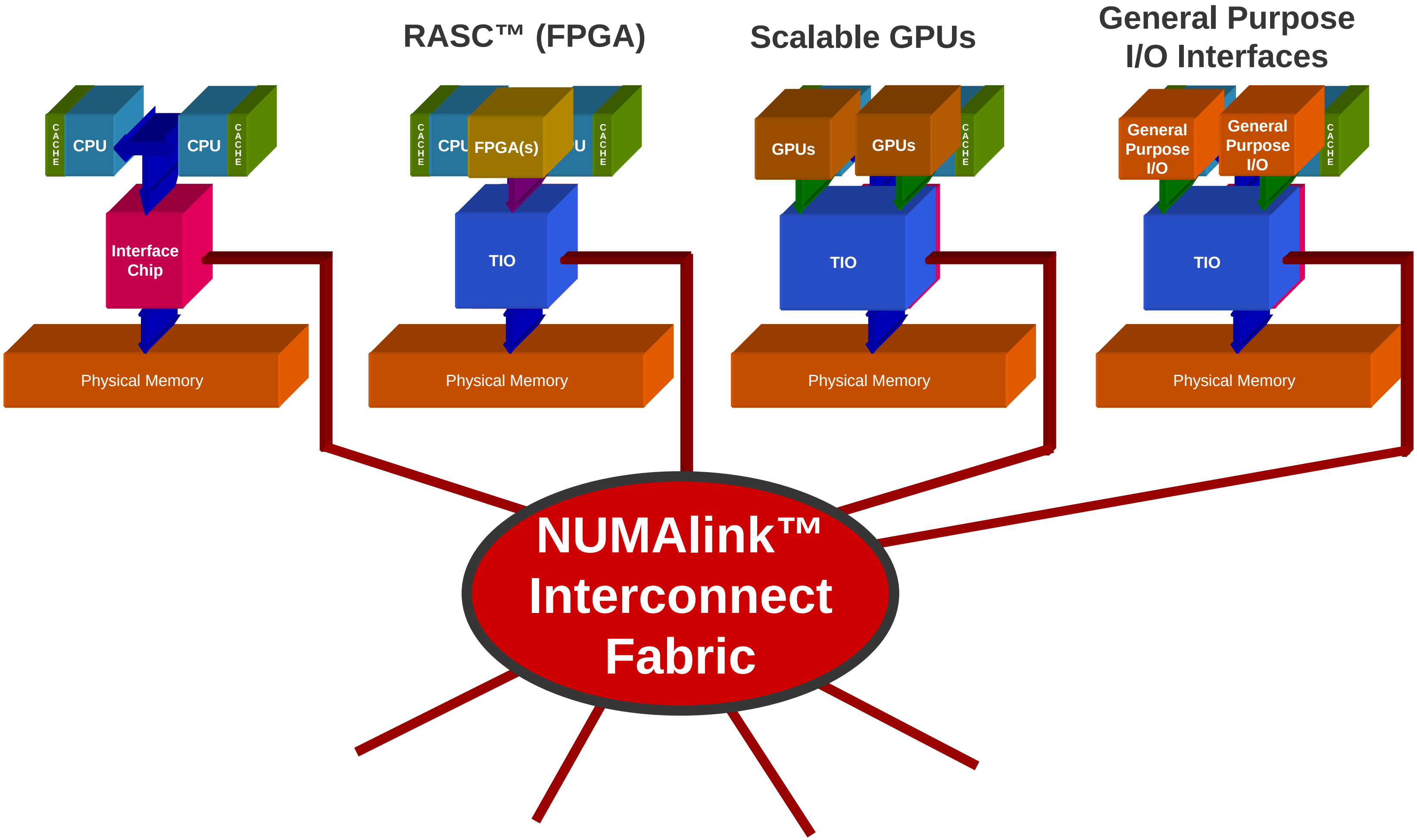
Peer I/O: Increased I/O Flexibility & Performance



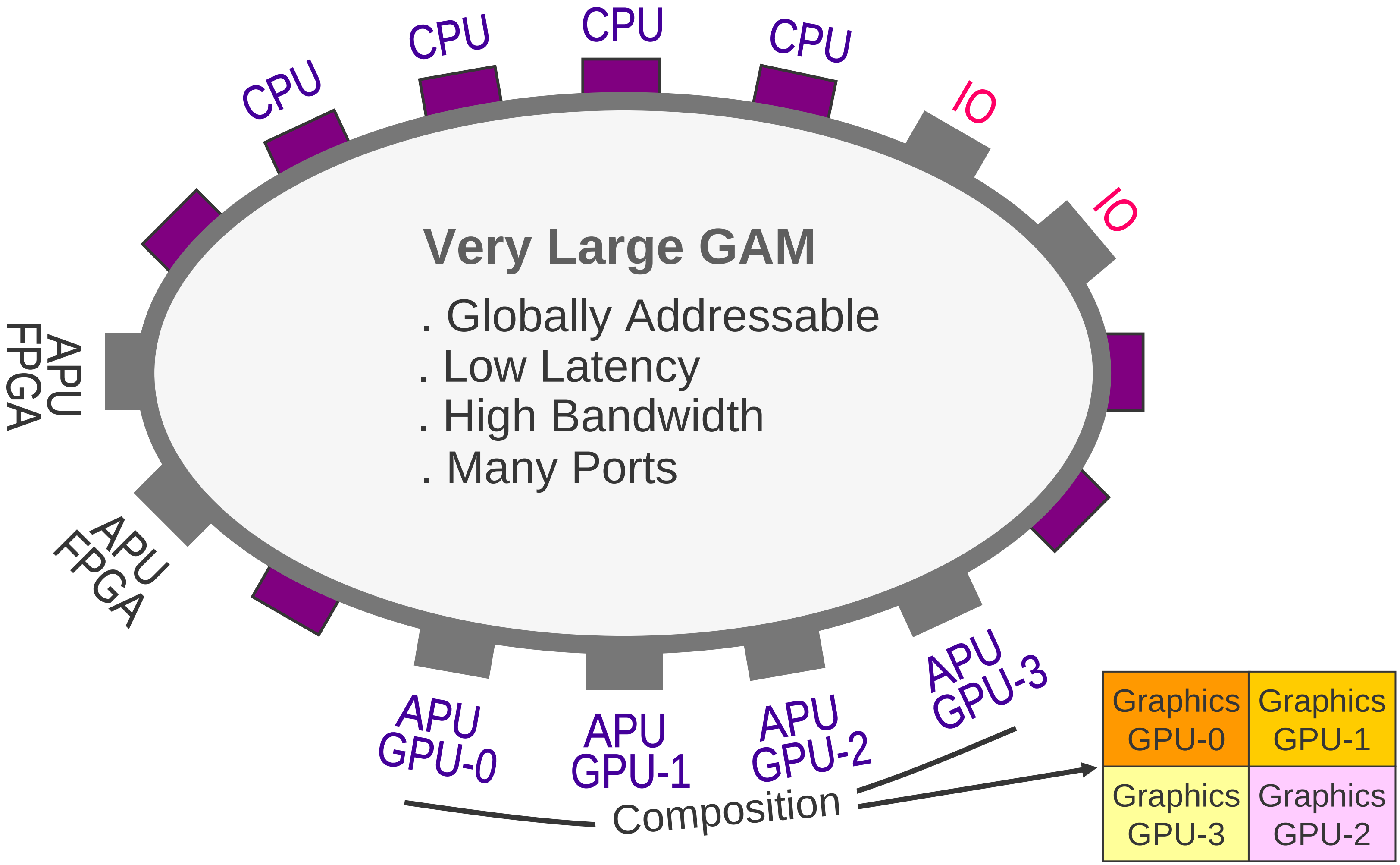


SGI Scalable ccNUMA Architecture

Multi-Paradigm Computing Architecture



Data-Centric Architecture

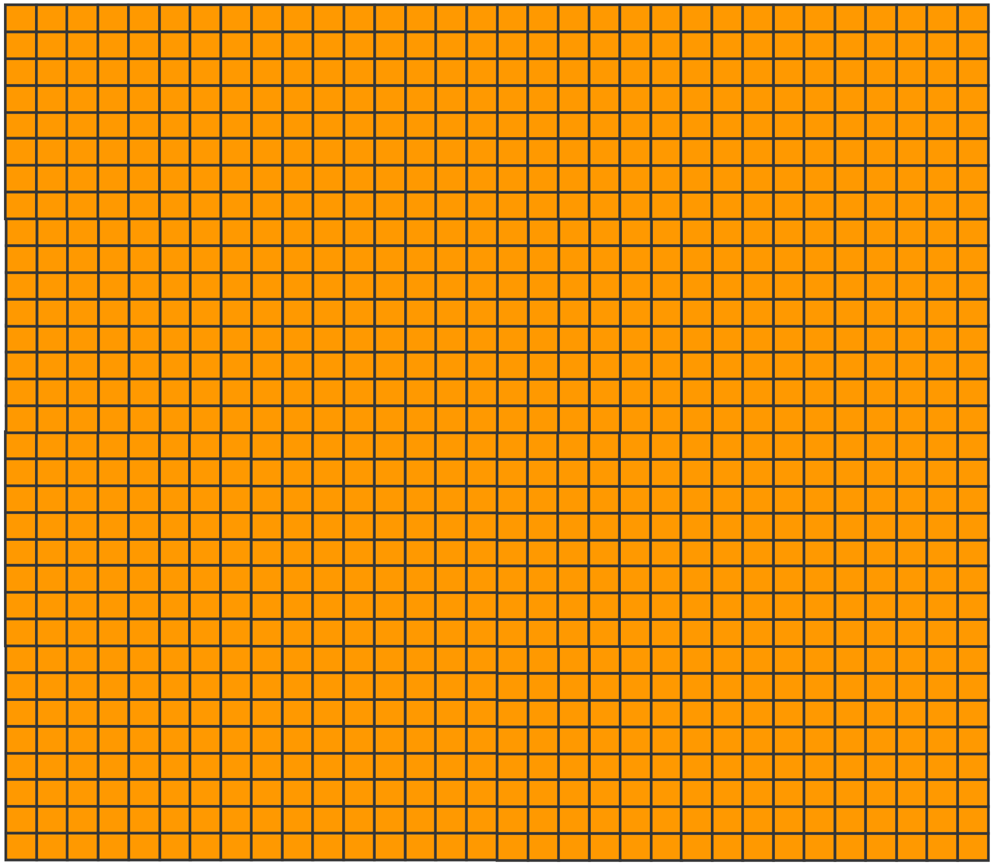


Very Large GAM

. Globally Addressable
. Low Latency
. High Bandwidth
. Many Ports

Graphics GPU-0	Graphics GPU-1
Graphics GPU-3	Graphics GPU-2

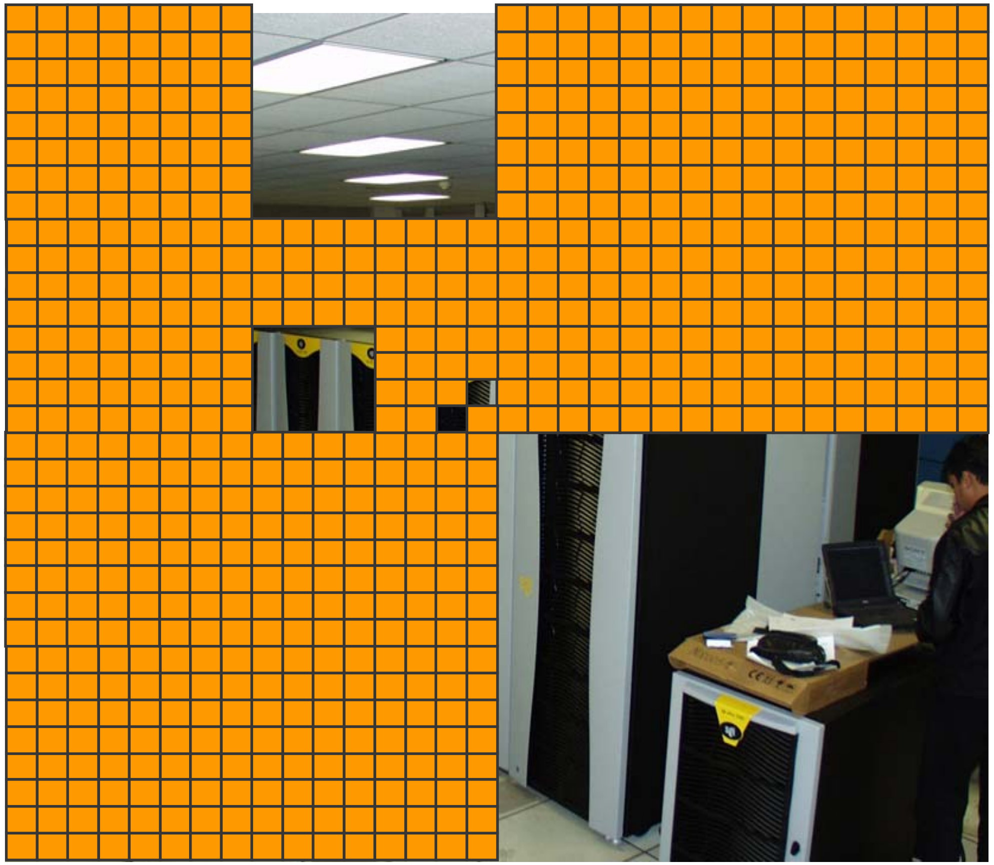
Big Data



1TB, 32*32=1024 elements

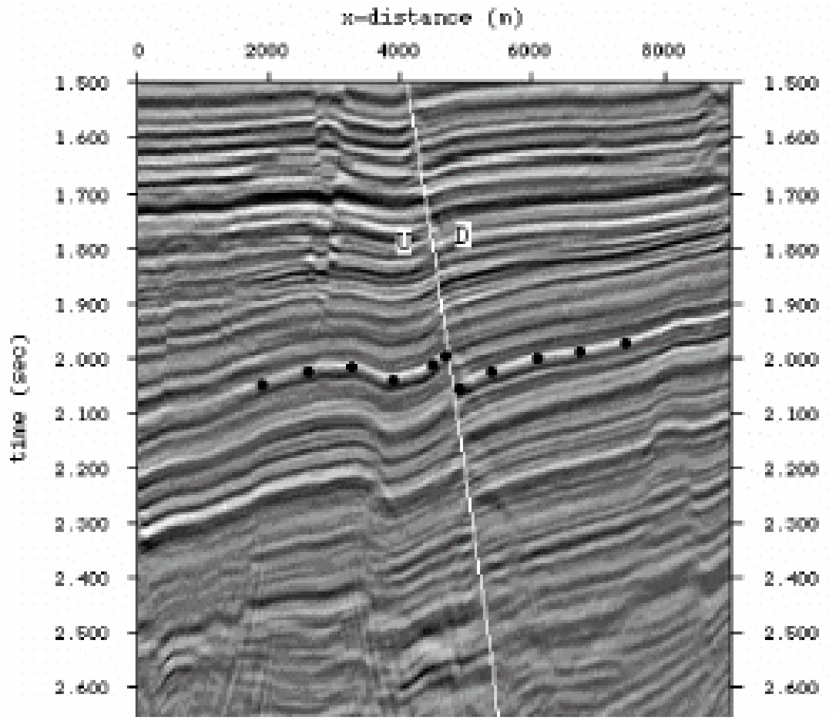
Each box represents 1GB

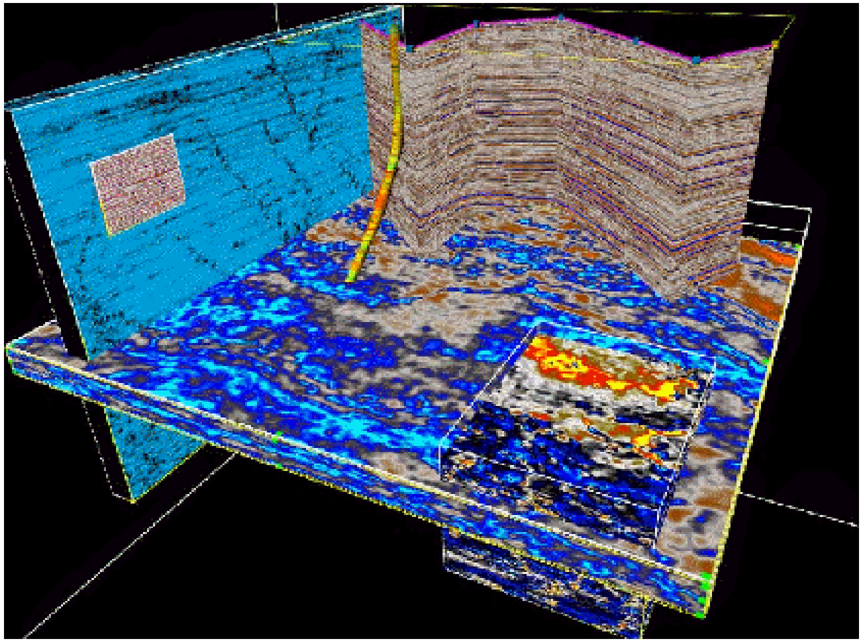
Big Data



Big Data

Big Datasets : 3D Interactive Visualization





1993
100 MB
10% viewed / year
~1 MB / month

40,000x Productivity →

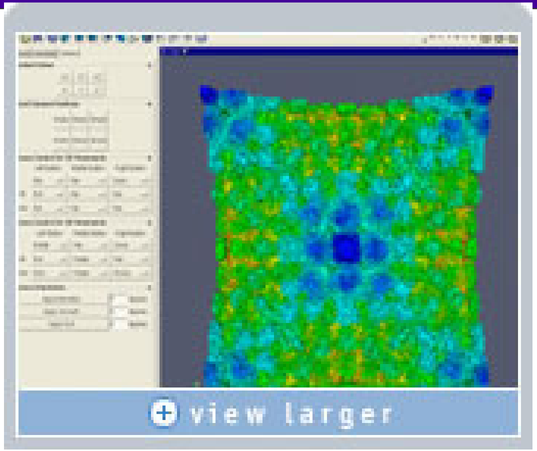
2004
400 GB
100% viewed / month
400 GB / month

Commodity GPU systems 5X the price of a Scale-up System

March 17, 2005

*n*VIDIA visualizes large data set

- •473 million triangles
- •128 GPU's on Dell Systems
- •~$1million system

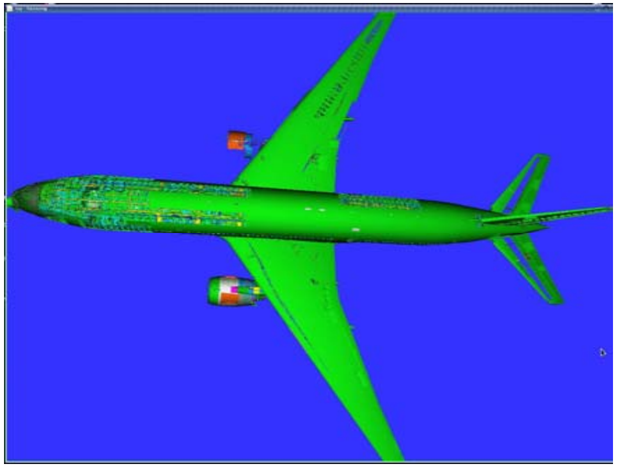


Compliments of nVIDIA

January 21, 2005

SGI visualizes large data set

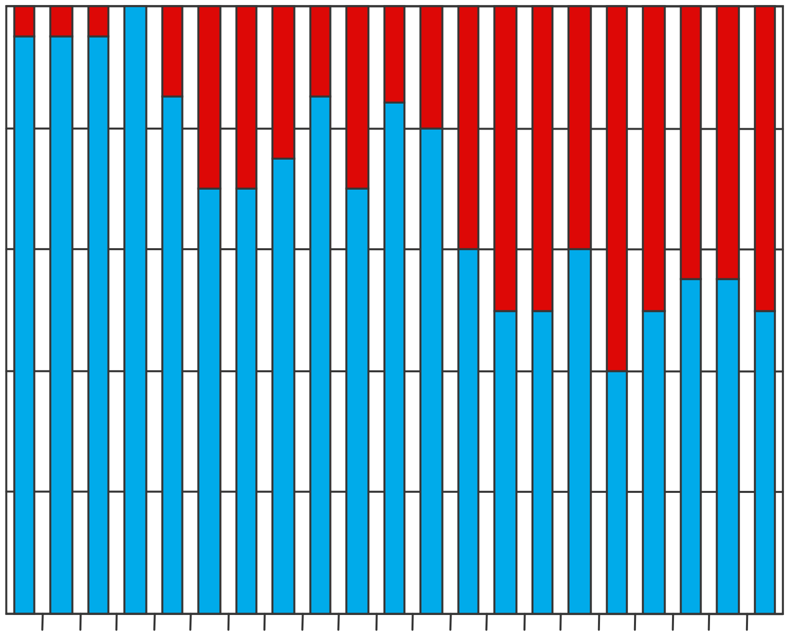
- •350 million triangles
- •12P, 56GB memory
- •Utilizes a ray tracer
- •~$180,000 system



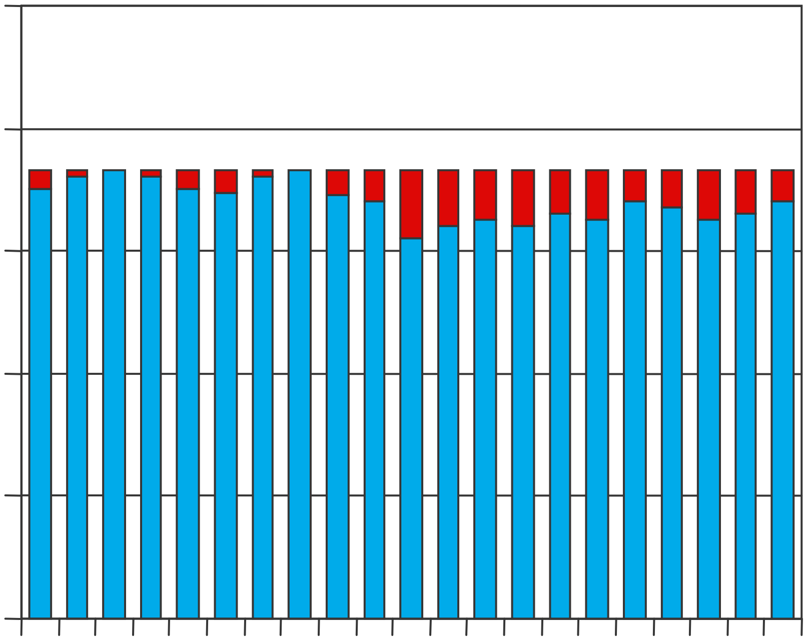
Compliments of Boeing

Dynamic Load Balancing

Load Balancing OFF



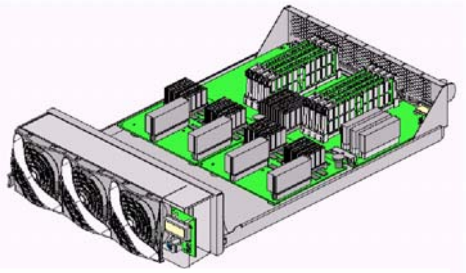
Load Balancing ON



Dimensions of Scalability

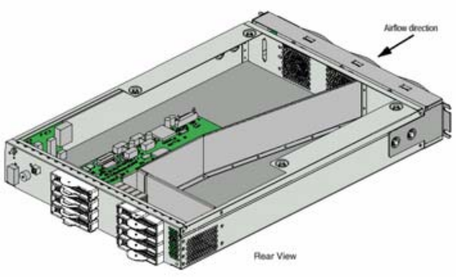
- Processors
- Processor bandwidth
- Memory bandwidth
- Memory capacity
- Interconnect bandwidth
- IO bandwidth
- Graphics processing
- Reconfigurable processing
- Other acceleration elements

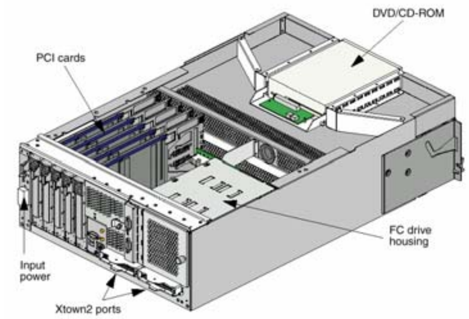
Origin3000 Building Blocks (Bricks)



C-brick
CPU Module

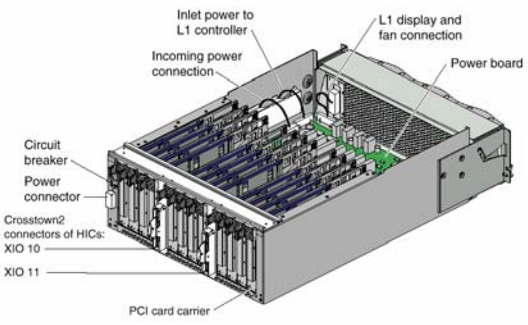
R-brick
Router Interconnect

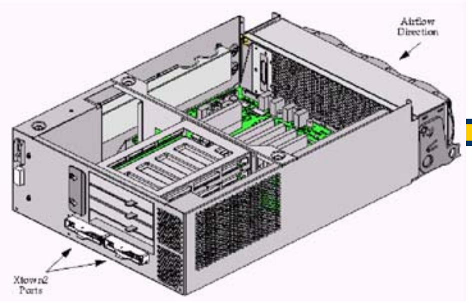




I-brick
Base I/O Module

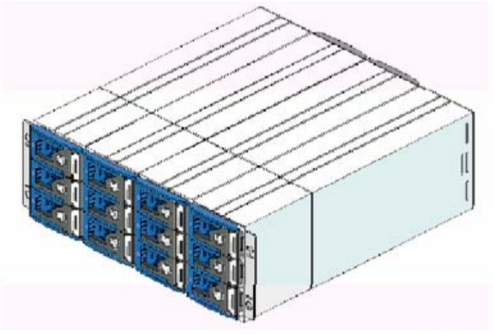
P-brick
PCI Expansion


Inlet power to L1 controller
L1 display and fan connection
Incoming power connection
Power board
Circuit breaker
Power connector
Crosstown2 connectors of HICs:
XIO 10
XIO 11
PCI card carrier

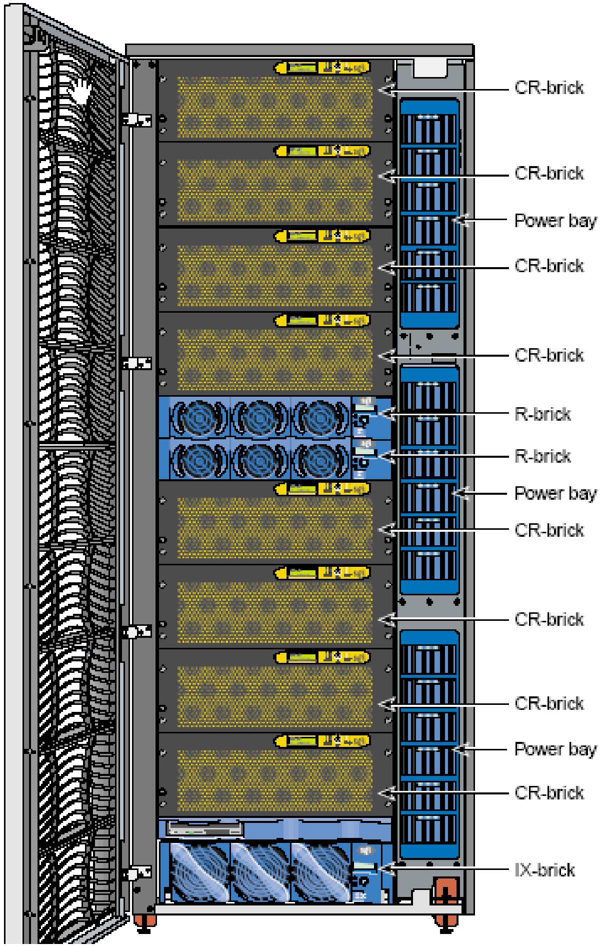


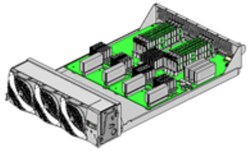
X-brick
XIO Expansion

D-brick
Disk Storage

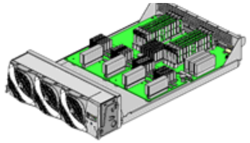


SGI Altix™ 3700 Bx2 Platform Introduction: Building Blocks

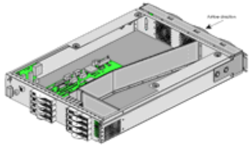




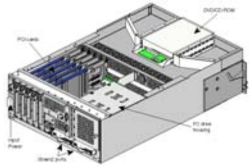
Itanium® 2 CR-brick
CPU and memory



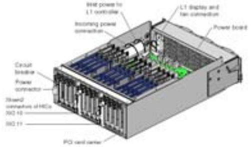
M-brick
Memory



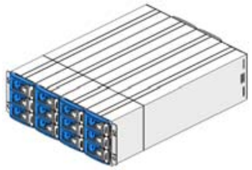
R-brick
Router interconnect



IX-brick
Base I/O module



PA-brick, PX-brick
PCI-X expansion



D-brick2
Disk expansion

SGI® Advanced Linux Environment With SGI ProPack

High-End Servers – Moving Forward: Altix® 4700 Platform….. Blade Packaging

- **Innovative Blade-to-NUMALink4 Concept:** Provides Unprecedented Versatility, Density

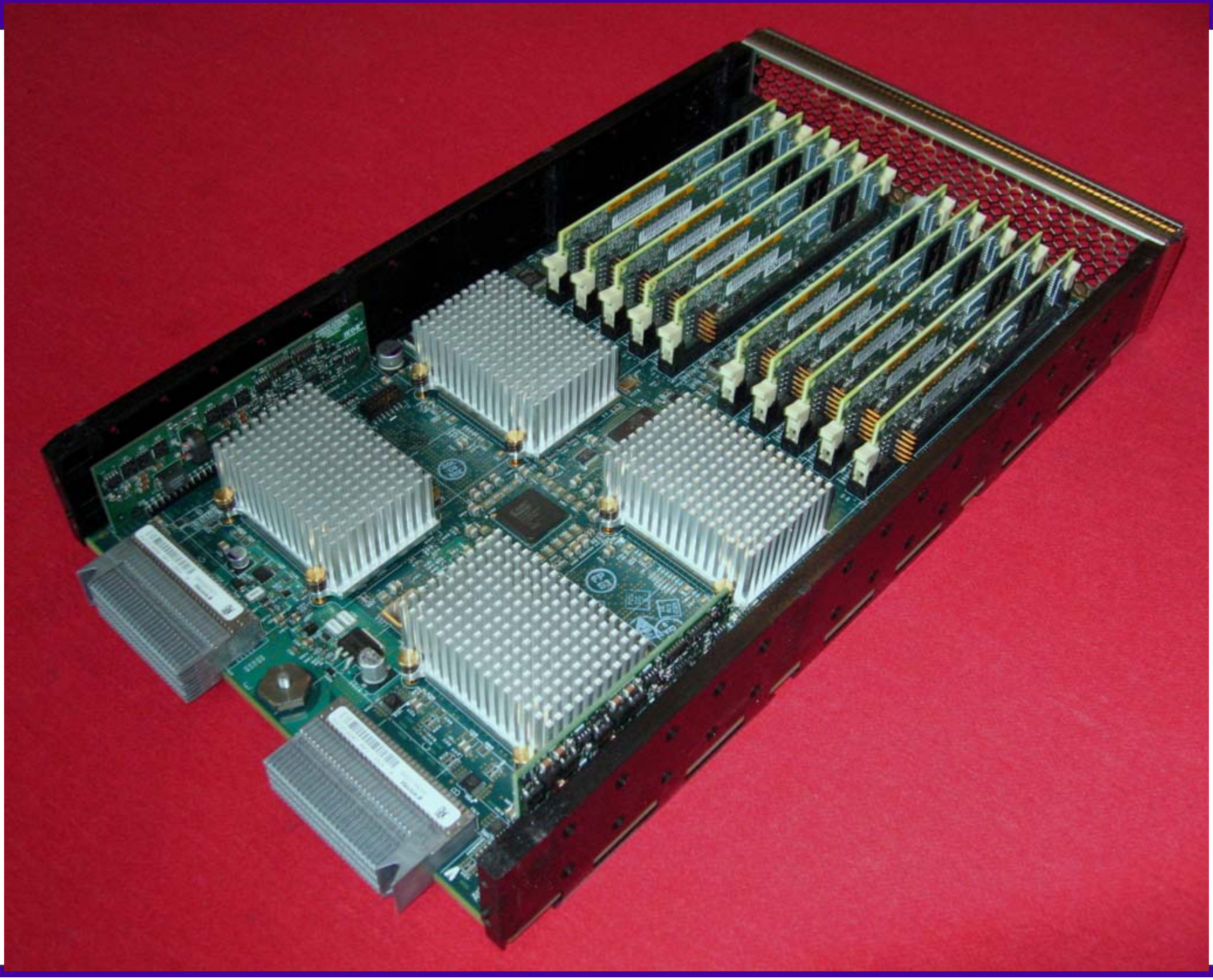
- **Blade Architecture Leads Next-Wave of HPC Blade-Based Platforms:** With Better Upgradeability, Expansion & Repair

- **Investment Protection:** Processor-Only Upgrade to Future Dual Core Processors

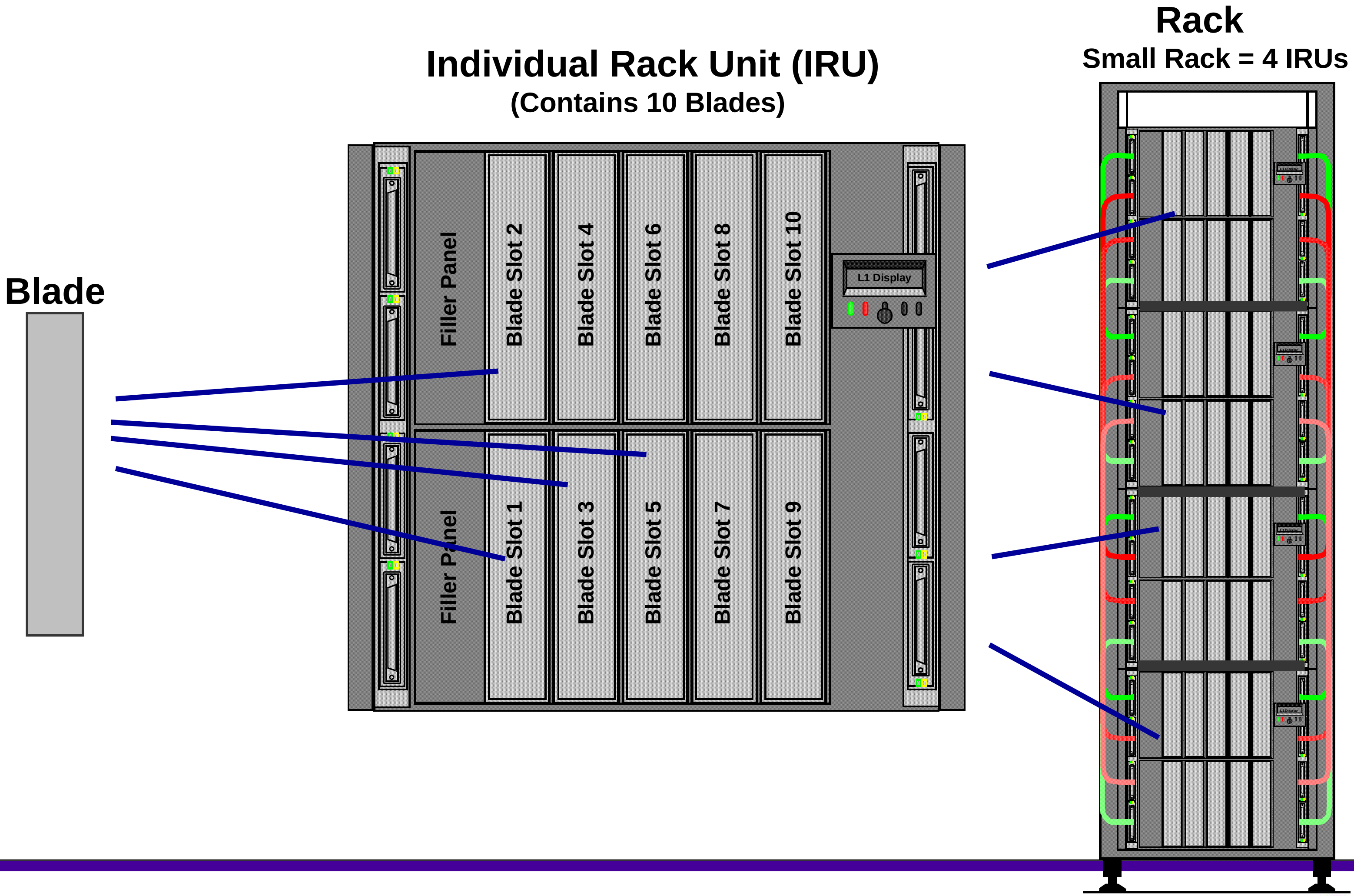
- **Enables Flexible Multi-Paradigm Computing:** Enhanced integrated RASC, Graphics

Next Generation RASC™ Technology

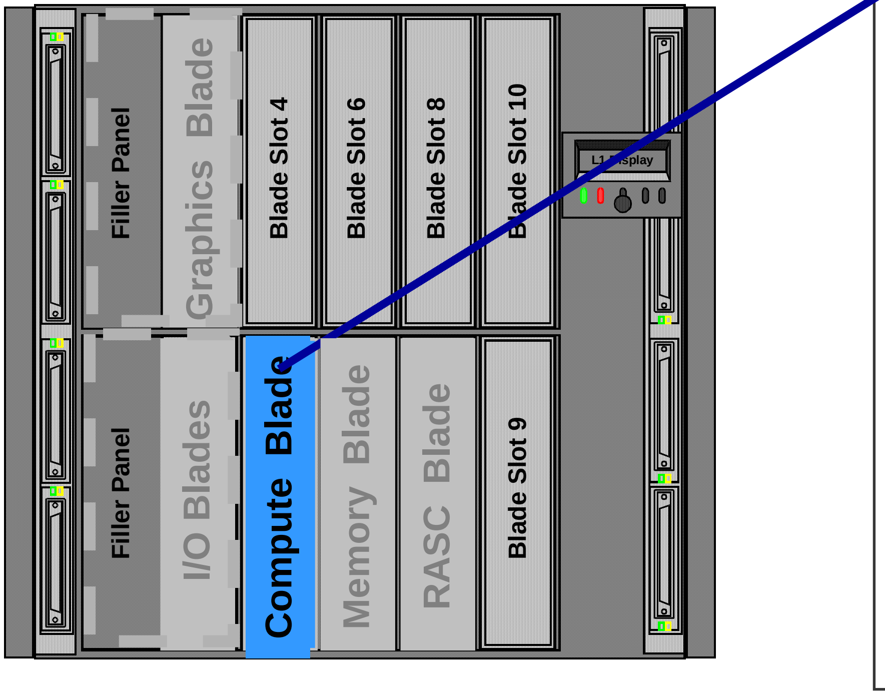
Blade Based Package



Standardized Blades, NUMAlink Backbone



Altix® 4700 Compute Blades

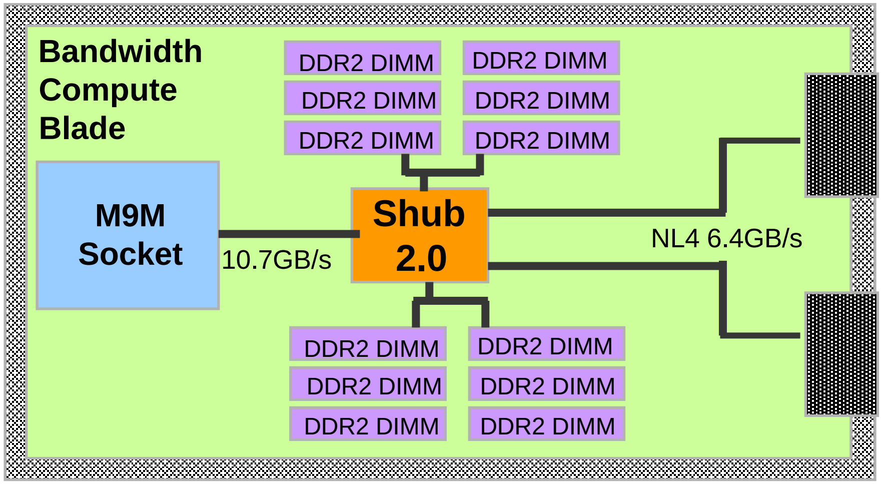


- Support for Madison9M Processors (Montecito/Montvale as Available)
- Two Compute Blade Options to Provide Different System Capabilities:
  - Best $/FLOP, Best Density (Density Compute Blade)

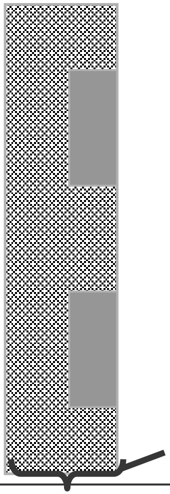
  OR
  - Best Performance, Memory BW (Bandwidth Compute Blade)

Altix® 4700 Compute Blades

Top View



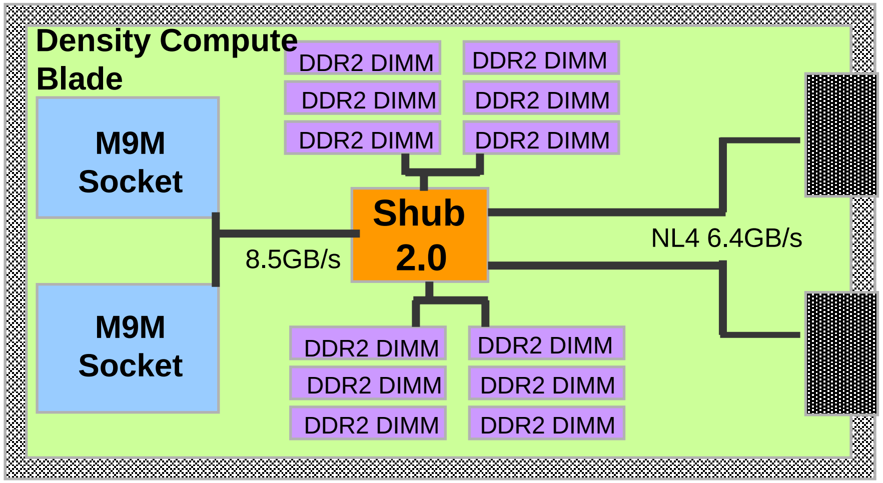
Front View



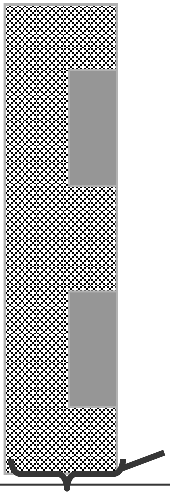
Highest Memory BW, Performance: Bandwidth Compute Blade

- 667MHz FSB Madison9M -> 10.7GB/s Local Memory Bandwidth
- 32 M9M Sockets / S-Rack
- Processors Supported: 1.66GHz/9M, 1.66GHz/6M Madison9M with 667MHz FSB
- Memory Sizes: 2G – 48G/core

Top View



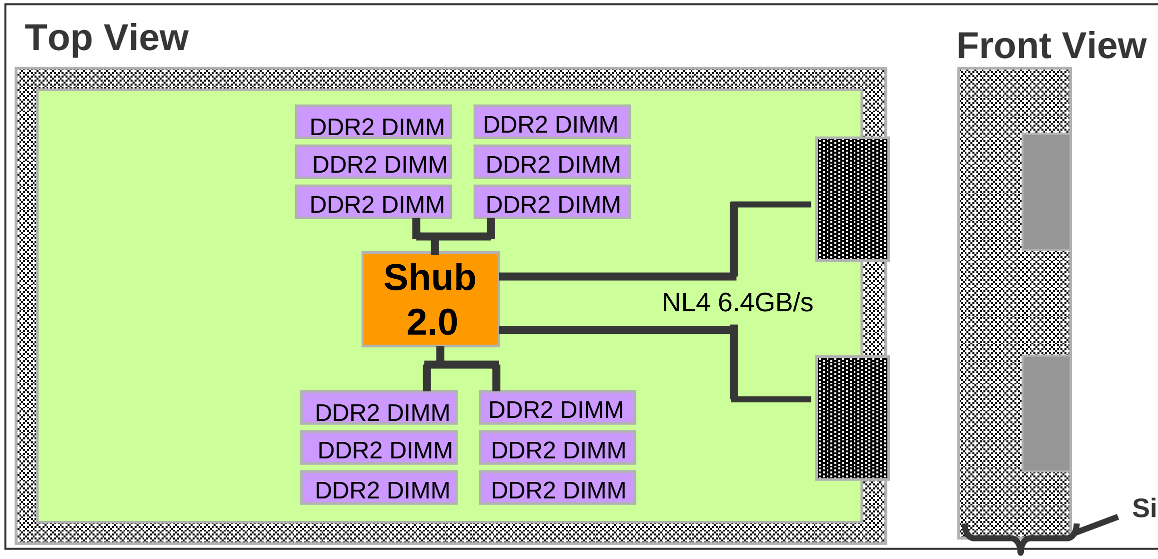
Front View



Best $/FLOP, Best Density: Density Compute Blade

- 533MHz FSB Madison9M -> 8.524GB/s Local Memory Bandwidth
- 64 M9M Sockets / S-Rack
- Processors Supported: 1.6GHz/9M, 1.6GHz/6M Madison9M with 533MHz FSB
- Memory Sizes: 1G – 24GB/core

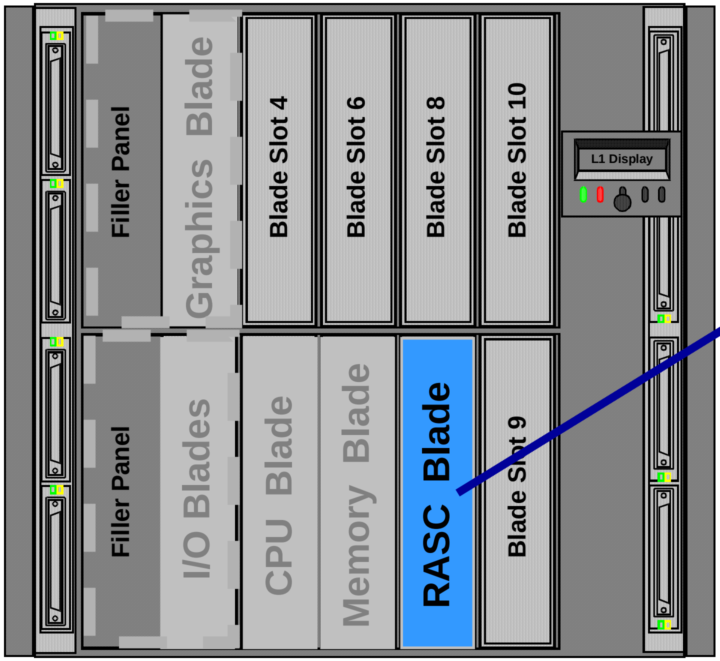
Memory Blade



Memory Blade:

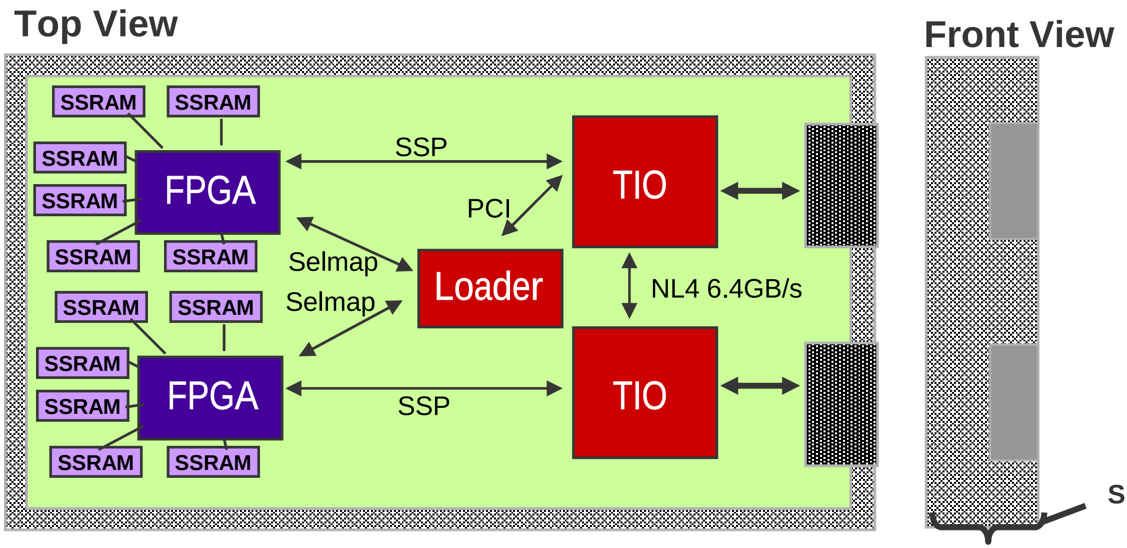
- Scale Memory Independently with 12 DDR2 DIMM Slots Per Blade
- Up to 128TB

Altix® 4700 RASC Blade



- RASC Blade
  - Abacus Computation Blade
  - Enhanced Performance, Tightly Integrated

RASC Blades – Cont.

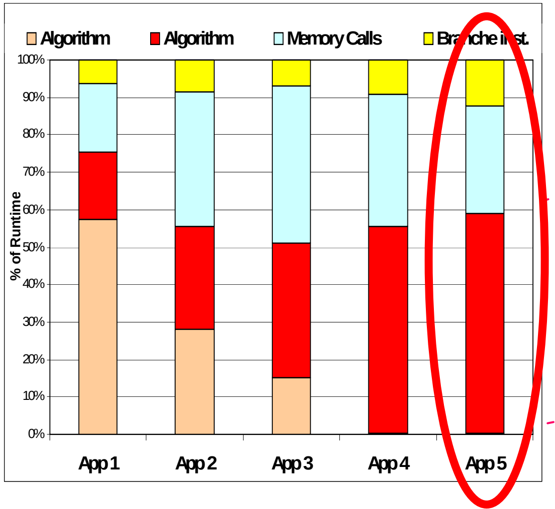


Abacus Computation Blade:

- New Levels of Performance:
  - High Performance V4LX160 FGPA with 160K Logic Cells
  - Increased Memory Sizes,12 DIMM per Blade
- Optional Brick Packaging for Legacy Platforms

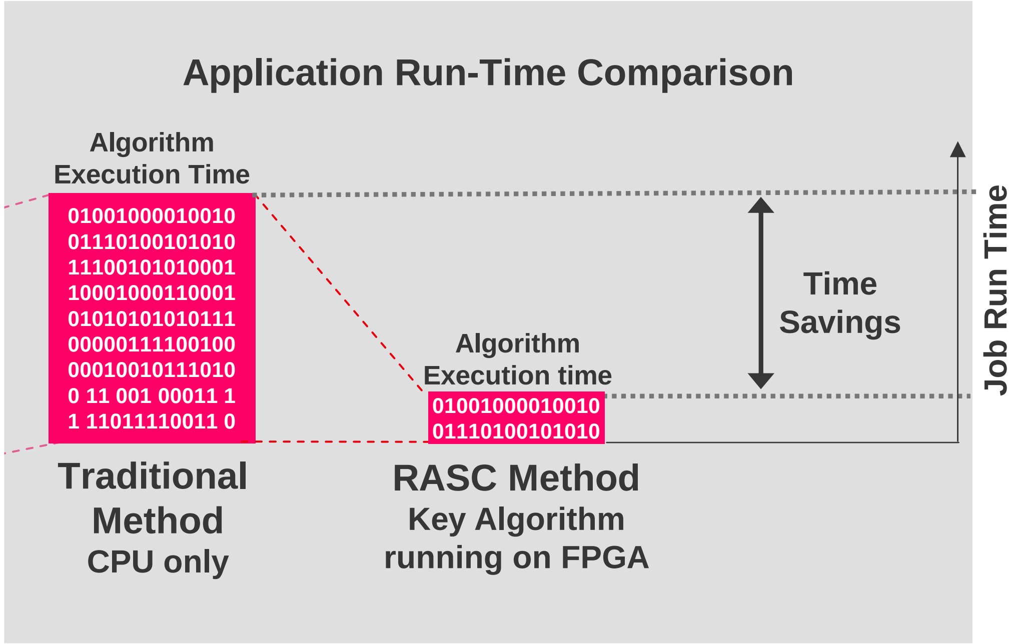
How does RASC™ Technology Differ from Traditional CPUs?

Compare Application Run Time %'s



Identify RASC appropriate algorithm

Export Algorithm to RASC



Directly map computationally-intensive algorithms to hardware with RASC

Application Segments

Application segments	Sample applications
Image and video processing	Transcoding (digital watermarks, format conversion), compression (JPEG, MPEG), color correction, ray-tracing, edge detection (Sobel)
Digital Signal Processing	FFT, IFFT, Filtering (FIR and IIR)
Network and Communication	Interleaver/de-interleaver, coding/decoding (Reed Solomon, Viterbi), convolution encoders, encryption, error correction, packet processing (IPsec)
Database Acceleration	Query, sorting, pattern recognition, data compression
HPC Algorithm Acceleration– Gov/Defense	MATLAB, STAR-P, random number generators, Sigint/Elint, image recognition (radar/vision/IR), DEM
HPC Algorithm Acceleration– Bioinformatics	Blast, Smith-Waterman

Ease of Use

- **Leverage 3rd Party Std Language Tools**
  - **Celoxica, Mitrionics, Starbridge Systems, Nallatech**

- Developed an FPGA aware version of GDB
  - Capable of debugging the FPGA and System Software
  - Capable of multiple CPUs and multiple FPGAs

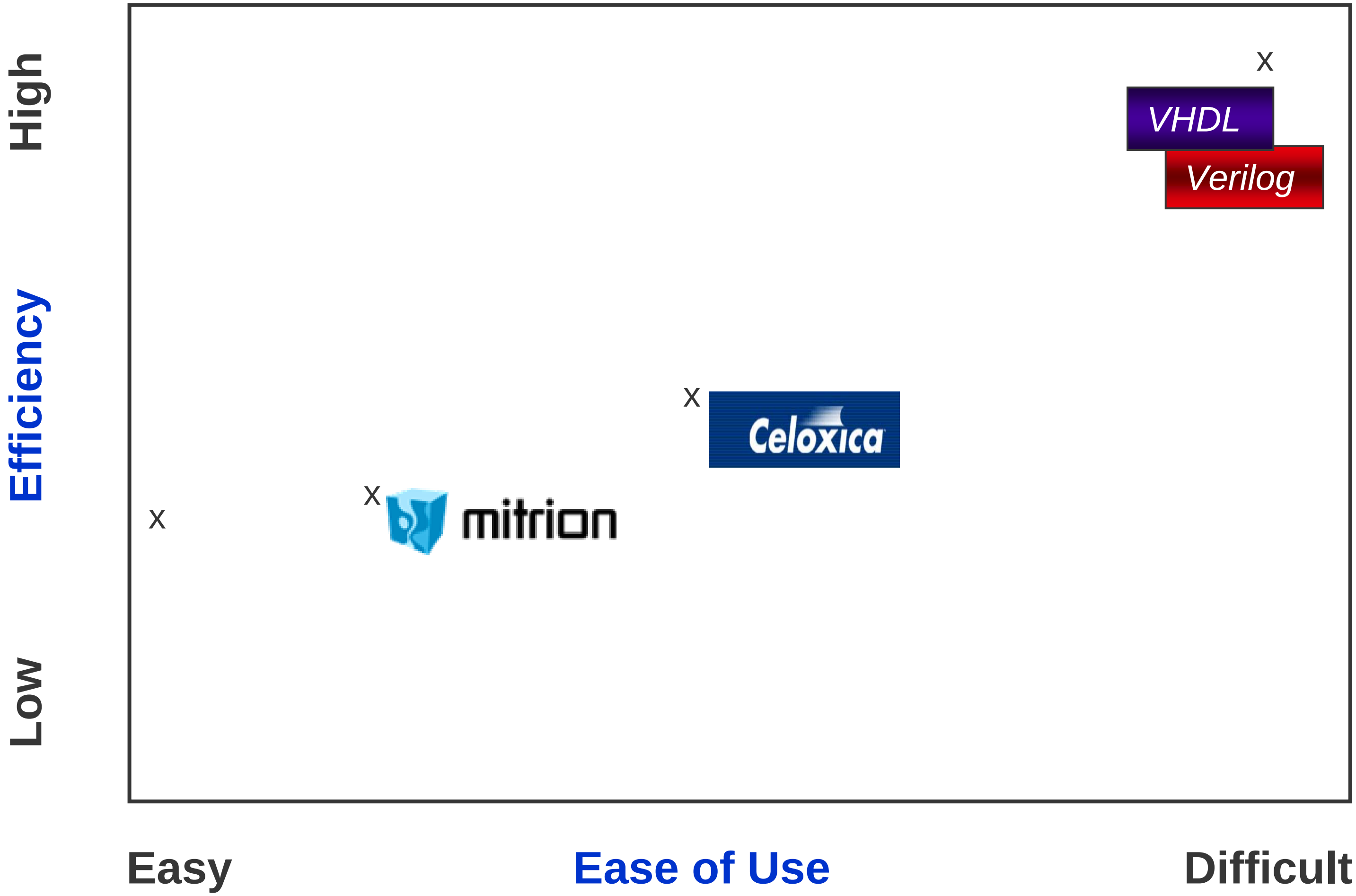
- Developed RASC Abstraction Layer (RASCAL)

- Provide for HDL modules
  - Integrated environment with debugger
  - Highest performance

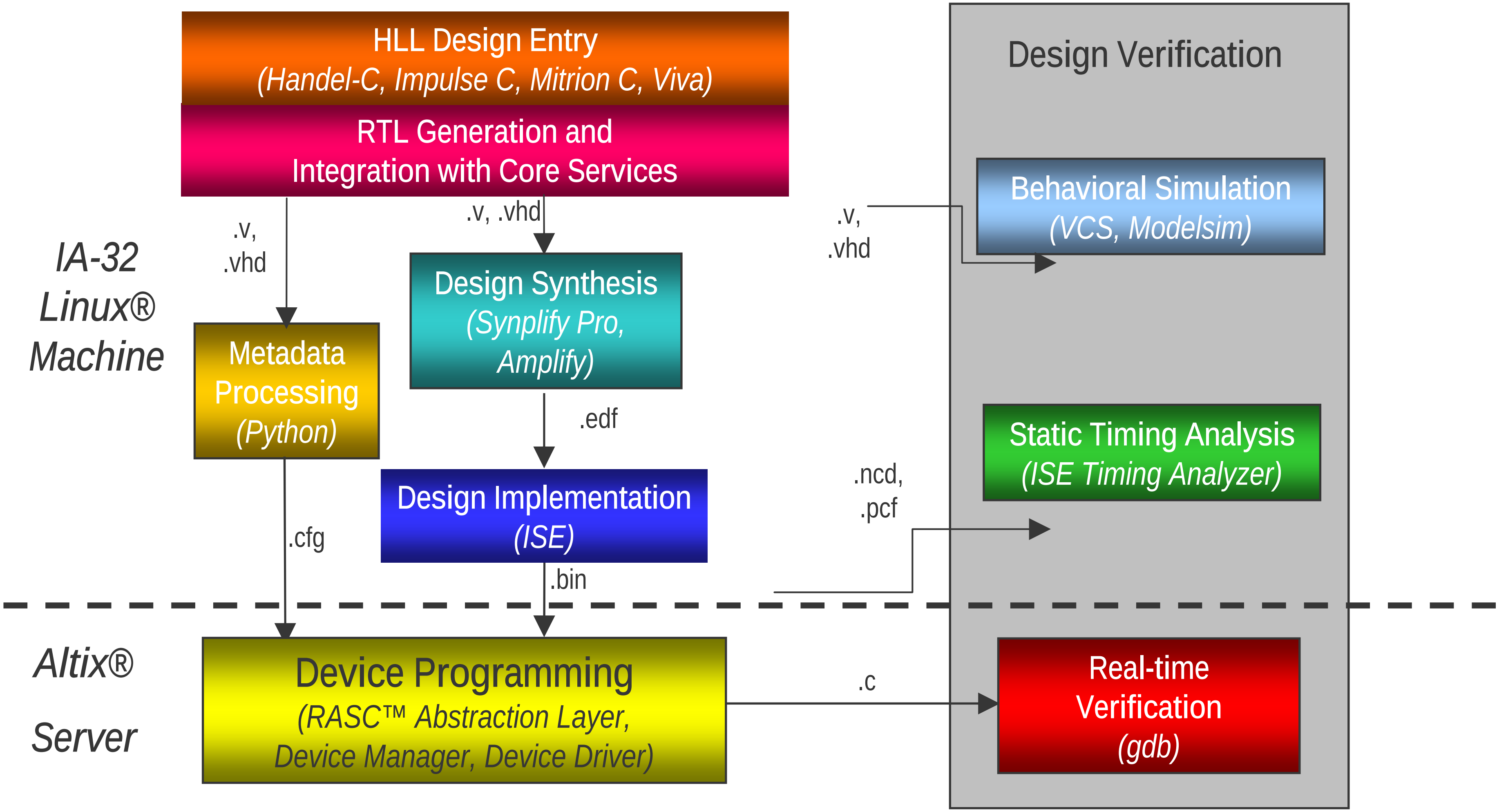
3rd Party Tools

- **Celoxica – http://www.celoxica.com**
  - **Handel-C**
- **Mitrionics - http://www.mitrionics.com**
  - **Mitrion C**
- **Starbridge Systems - http://www.starbridgesystems.com/**
  - **Viva graphical development environment**
- **Nallatech - http://www.nallatech.com/**
  - **SGI strategic partner**

Ease of Use v. Efficiency



Bitstream Generation… HLL Tools



Ease of Use

- Leverage 3rd Party Std Language Tools
  - Celoxica, Impulse Acceleration, Mitrion, Starbridge Viva
  - In discussions with other HLL tool vendors

- **Developed an FPGA aware version of GDB**
  - **Capable of debugging the FPGA and System Software**
  - **Capable of multiple CPUs and multiple FPGAs**

- Developed RASC Abstraction Layer (RASCAL)

- Provide for HDL modules
  - Integrated environment with debugger
  - Highest performance

FPGA Aware Debugger

- Based on Open Source GNU Debugger (GDB)
- Uses extensions to current command set
- Can debug host application and FPGA
- Provides notification when FPGA starts or stops
- Supplies information on FPGA characteristics
- Can “single-step” or “run N steps” of the algorithm
- Can HLL line step per C-line source
- Dumps data regarding the set of “registers” that are visible when the FPGA is active

GDB Debugging Environment

```
(gdb) fpgastep

(gdb) p/x  $a
$6 = 0x444433

(gdb) p/x $b
$7 = 0x111122

(gdb) p/x $tmp
$8 = 0x555533

(gdb) fpgastep

(gdb) p/x $tmp
$9 = 0x555533

(gdb) p/x $c
$10 = 0x331222

(gdb) p/x $d
$11 = 0x111022
```

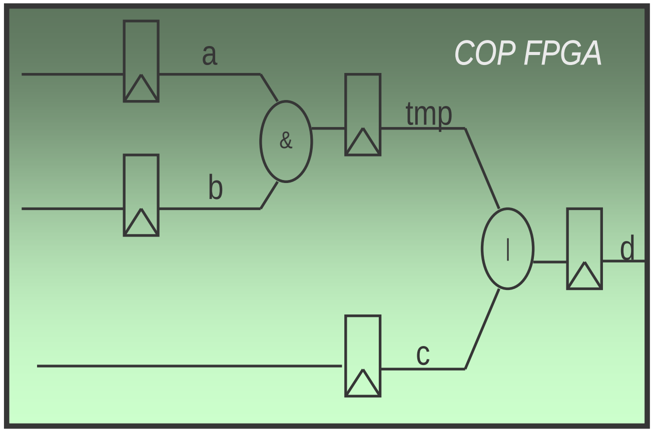
Debugger running

in real time

Algorithm.c

```
tmp = a & b;

d = tmp | c;
```



Ease of Use

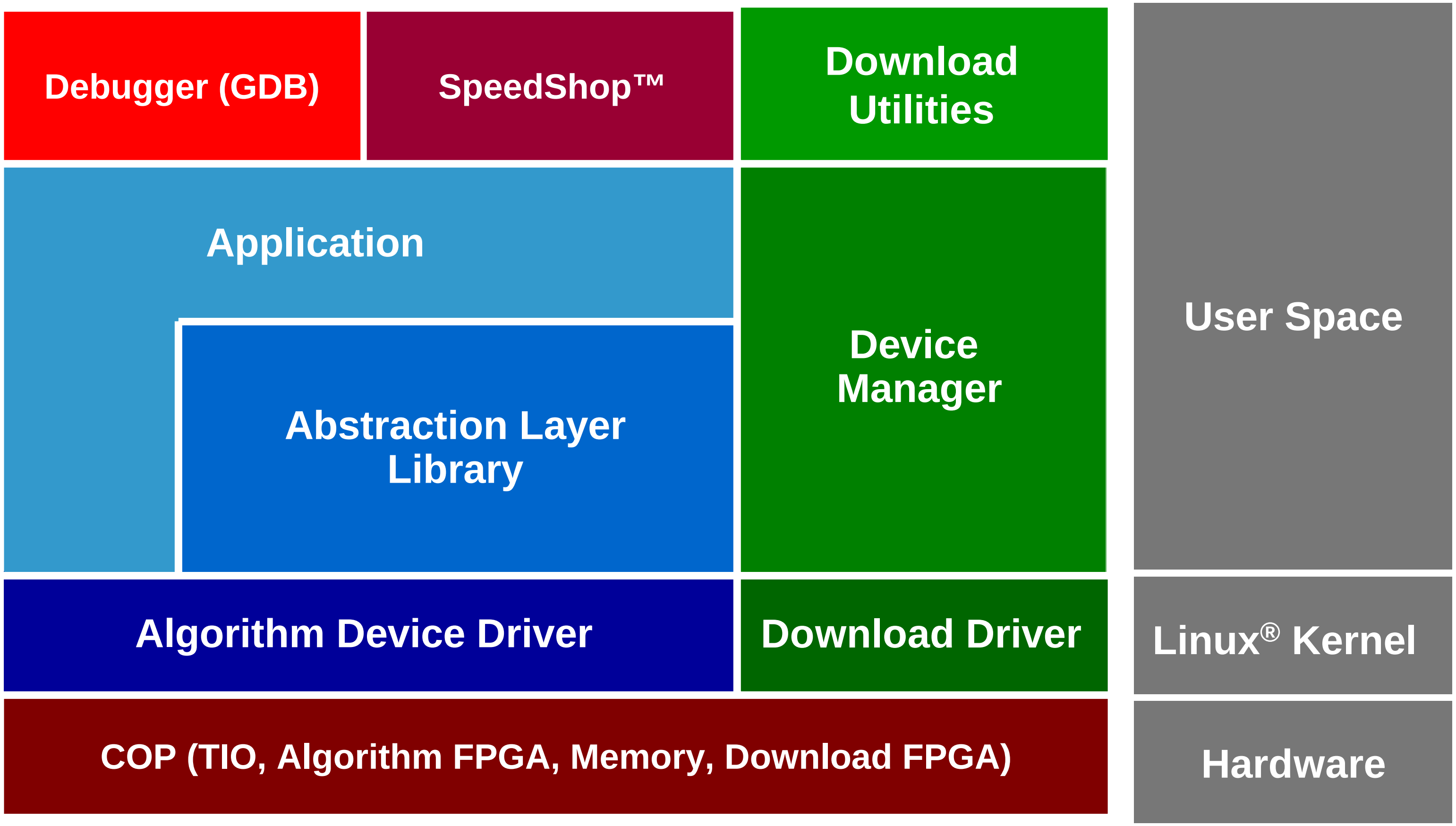
- Leverage 3rd Party Std Language Tools
  - Celoxica, Impulse Acceleration, Mitrion, Starbridge Viva
  - In discussions with other HLL tool vendors

- Developed an FPGA aware version of GDB
  - Capable of debugging the FPGA and System Software
  - Capable of multiple CPUs and multiple FPGAs

- **Developed RASC Abstraction Layer (RASCAL)**

- Provide for HDL modules
  - Integrated environment with debugger
  - Highest performance

RASC™ Software Stack



Debugger (GDB)

SpeedShop™

Download Utilities

Application

Abstraction Layer Library

Device Manager

User Space

Algorithm Device Driver

Download Driver

Linux® Kernel

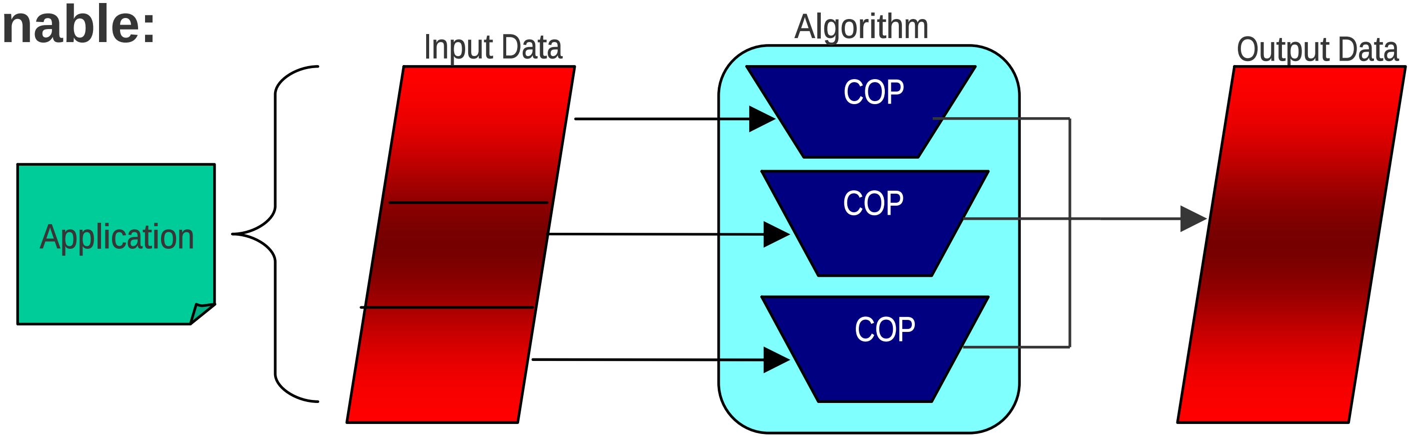
COP (TIO, Algorithm FPGA, Memory, Download FPGA)

Hardware

Abstraction Layer: Algorithm API

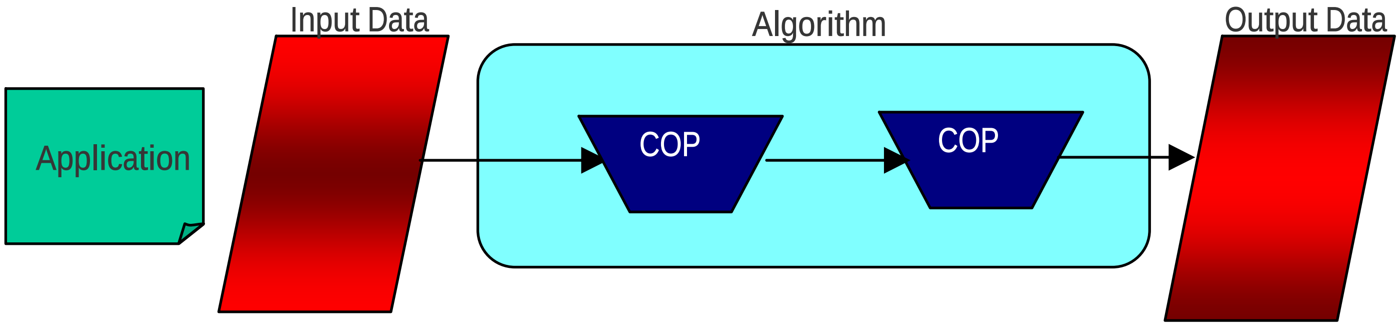
The Abstraction Layer’s algorithm API mirrors the COP API with a few additions that enable:

Wide Scaling



- and -

Deep Scaling



Working with industry/customers (www.openfpga.org) on API stds…

Ease of Use

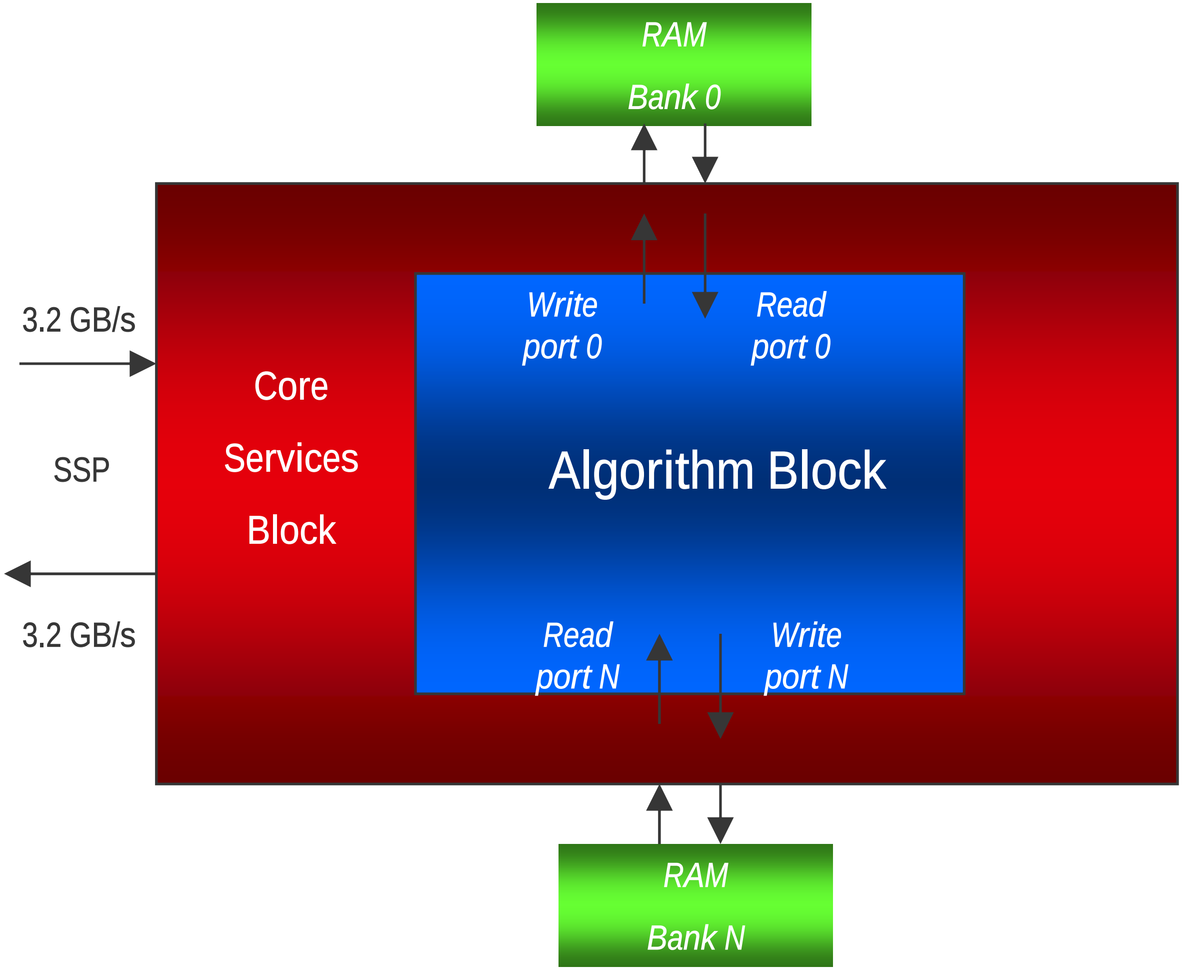
- Leverage 3rd Party Std Language Tools
  - Celoxica, Impulse Acceleration, Mitrion, Starbridge Viva
  - In discussions with other HLL tool vendors

- Developed an FPGA aware version of GDB
  - Capable of debugging the FPGA and System Software
  - Capable of multiple CPUs and multiple FPGAs

- Developed RASC Abstraction Layer (RASCAL)

- **Provide for HDL modules**
  - **Integrated environment with debugger**
  - **Highest performance**

FPGA Architecture Overview



RAM

Bank 0

3.2 GB/s

SSP

3.2 GB/s

Core

Services

Block

Write

port 0

Read

port 0

Algorithm Block

Read

port N

Write

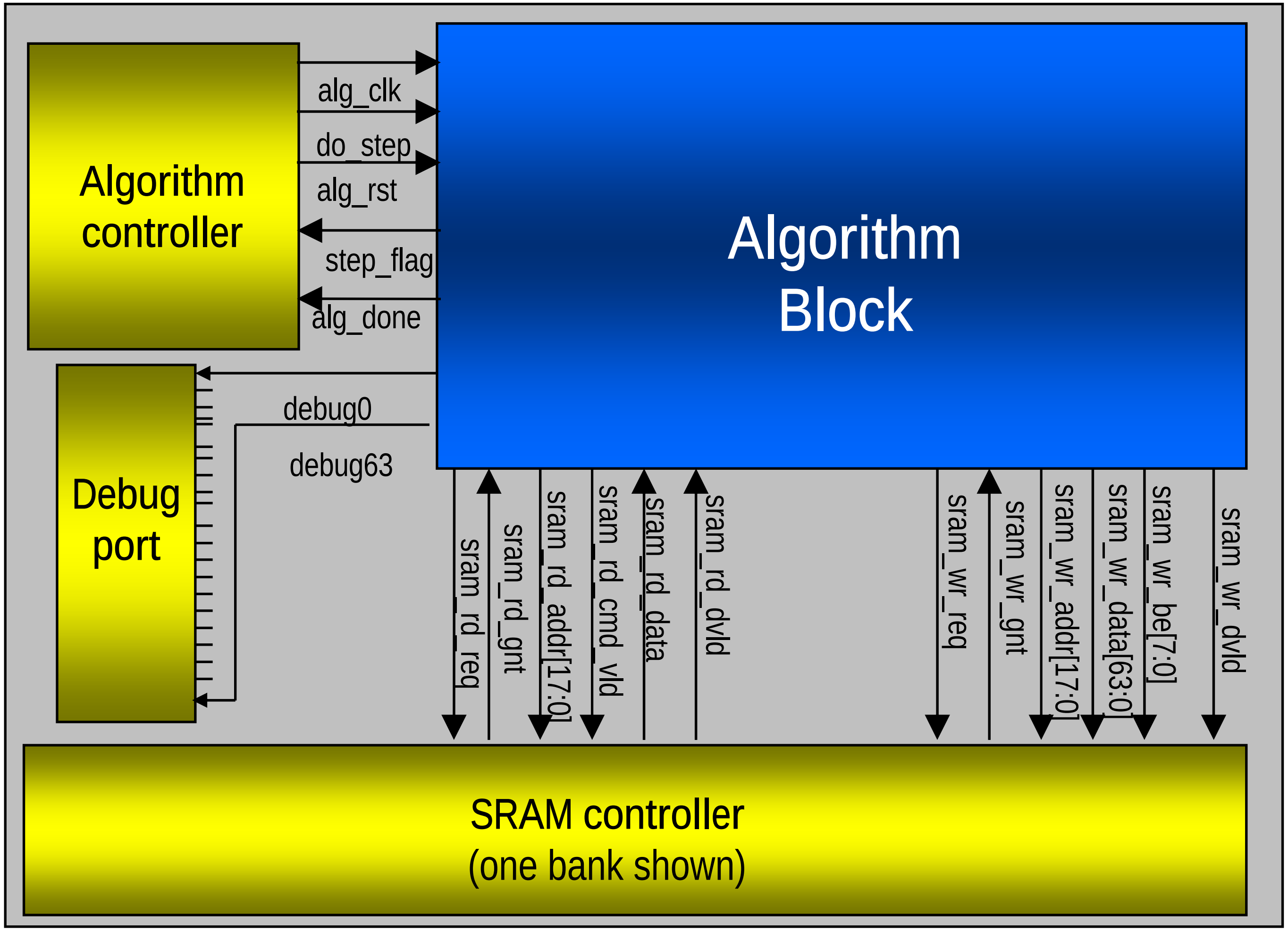
port N

RAM

Bank N

SGI Proprietary

Algorithm Block as Submodule



Verilog / VHDL Module Support

- Templates for Verilog and VHDL
  - Fast start to algorithm coding
- Provide a system simulation stub
  - Allows both simulation debug or system debug
- Provide source code for core service
  - Allows user to modify to meet special needs
- Extractor tools supports GDB meta-data
  - Application and FPGA debugging

RASC™ Technology — Demonstrated Application Speed-up

Bit Manipulation (Cryptography)[1]

- 79x 1.5GHz Intel® Itanium® 2 Processor (single RASC Unit)
- 119x 1.5GHz Intel® Itanium® 2 Processor (dual RASC Unit)

Graphics Edge Detection[1]

- 7.4x 1.5GHz Intel® Itanium® 2 Processor (single RASC Unit)

Customer Application

- 20,000x speedup on scalar microprocessor

- **EXERGY – MAPLD 2005 paper 190**

[1] Based on internal testing

RASC Platform Capabilities

- **Direct Connection to NUMAlink4**

  6.4GB/s/connection

- **Fast System Level Reprogramming of FPGA**

  FPGA load at memory speeds

- **Atomic Memory Operations**

  Same set as System CPUs

- **Hardware Barriers**

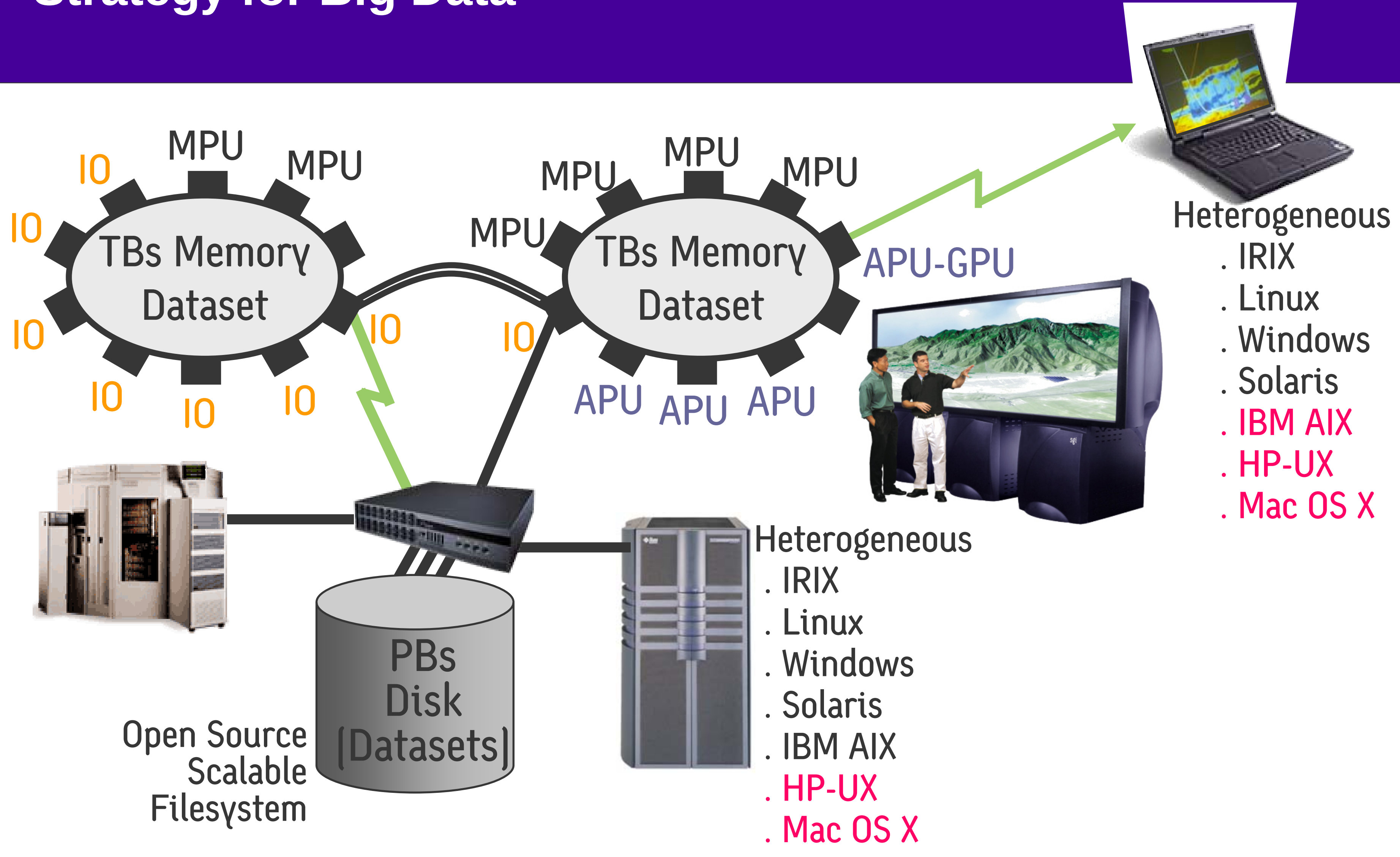
  Dynamic Load Balancing

- **Configurations to 8191 NUMA/FPGA Nodes**

  Scalability

Thank You

Strategy for Big Data



MPU MPU MPU IO IO IO IO IO IO

TBs Memory Dataset

MPU MPU MPU MPU

TBs Memory Dataset

IO IO

APU-GPU

APU APU APU

PBs Disk (Datasets)

Open Source Scalable Filesystem

Heterogeneous
. IRIX
. Linux
. Windows
. Solaris
. IBM AIX
. HP-UX
. Mac OS X

Heterogeneous
. IRIX
. Linux
. Windows
. Solaris
. IBM AIX
. HP-UX
. Mac OS X

SGI® RASC™ Technology Summary

- **Tightly coupled, high bandwidth/low latency integration into NUMA fabric**
  - Significant bandwidth advantage (6.4GB/s)
  - Coherent shared memory access
  - Atomic memory operations
  - Scalability (wide scaling and deep scaling)
- **Orders-of-magnitude performance improvement and application speedup**
  - Beneficial when running data intensive applications critical to oil and gas exploration, defense and intelligence, bioinformatics, medical imaging, broadcast media, and other data-dependent industries.
- **Ease of programming—complete software stack**
  - RASClib (API and core services library) provides abstraction layer to support reconfigurable elements in a multi-processing, multi-user environment
  - Fully integrated third-party party HLL development tools
  - FPGA-aware enhancements to GNU debugger (open-source)
- **Add-in module that seamlessly operates with SGI® Altix® servers and Silicon Graphics Prism™ visualization systems**

Multi-Paradigm Computing
Other Non-traditional Processing Initiatives

- **GPU-based processing**
  - High potential performance (200-300GF peak today) and performance/price on single precision floating point applications…clear roadmap to future semiconductor process technologies
  - SGI working with SI on scaling to multiple GPUs and on development environment/programming paradigms…initial focus on signal processing apps

- **Specialized processors… ClearSpeed™ processors, custom processors (MD-GRAPE, classified chip)**
  - High potential performance/watt on certain apps

This slide contains forward-looking statements. The results and forecasts as stated may vary. Other risks and uncertainties relating to this slide may be found in the "Safe-Harbor" statement at the beginning of this presentation.