

SMP, Workstation Cluster to MPPs: Evaluation for ECS Science Algorithms

Narayan Prasad

Marek Chmielowski

Scott Bramhall

726-PP-001-001

April 7, 1995

NP-1

Roadmap



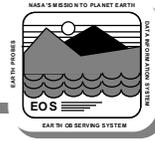
- Objectives
- Algorithm overview
- ECS STL Prototyping on SMP and Workstation cluster
- Prototyping on MPPs at HPCC sites (JPL and NASA/Ames)
- Comparison of MPPs with SMP and workstation cluster
- Conclusions

Note: Appendix contains additional slides for reference

726-PP-001-001

NP-2

Objectives

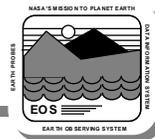


- Demonstrate migration of parallel applications developed on SMP and workstation cluster to MPPs
- Study MPPs from a new perspective for Release B and beyond using HPCC resources at JPL and NASA/Ames
- Evaluate parallelization compilers on MPPs
- Compare performance of SMP, workstation cluster and MPP to determine suitability of applications
- Determine and assess workload criteria for maximizing performance of parallel programs on MPPs
- Provide input and lessons learned to science algorithm developers

726-PP-001-001

NP-3

Algorithm Overview



- Algorithm
 - SeaWinds: ECS heritage, Level 2 (retrieve wind vectors over oceans from backscattered microwave radiation)
- Input data
 - simulated backscattered radiation measurements of 25 km x 25 km cells on ocean surface (1/3 of a single revolution)
 - Digital Weather Model (DWM) data
 - Backscattering model tabulated functions & other parameter data
- Processing
 - initialization (read DWM data, tabulated model functions, etc.)
 - wind vector retrieval (geolocate and group cells, compute wind parameters for each cell, remove ambiguities, etc.)
 - output product generation
- Output data
 - retrieved wind data, geolocated and annotated with DWM data

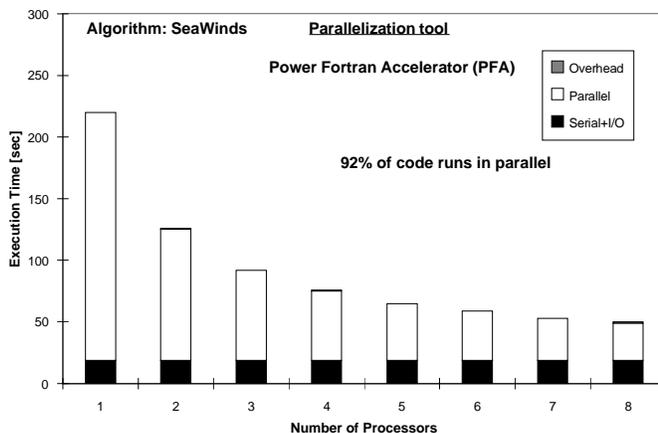
726-PP-001-001

NP-4

SeaWinds Execution Time - SGI Challenge XL



Execution Time - SGI Challenge XL;
Parallel Workload = 3.2 Giga Floating Point Operations



726-PP-001-001

Steps to parallelization

1. Performance analysis of serial program
2. Automatic parallelization
3. Analysis of PFA listing
4. Identify parallelization inhibitors
5. Interactive parallelization strategy**
 - identify parallelism

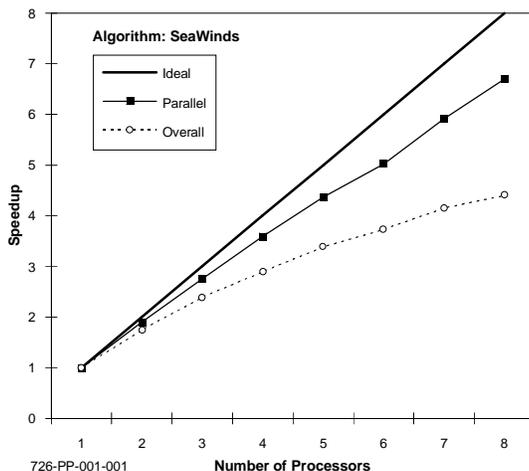
** Parallelization strategy to be developed at design

NP-5

Speedup on SGI Challenge XL



Parallel Workload = 3.2 Giga Floating Point Ops.



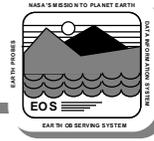
726-PP-001-001

Parallelization tool

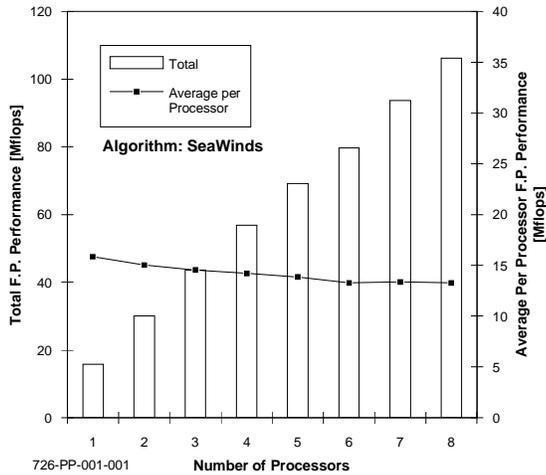
Power Fortran Accelerator (PFA)

NP-6

Floating Point Performance - SGI Challenge XL



Parallel Workload = 3.2 Gflop



Rated Peak MFLOPS = 75 per processor

Actual processor performance : ~20% of peak

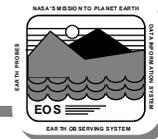
Parallelization tool

Power Fortran Accelerator (PFA)

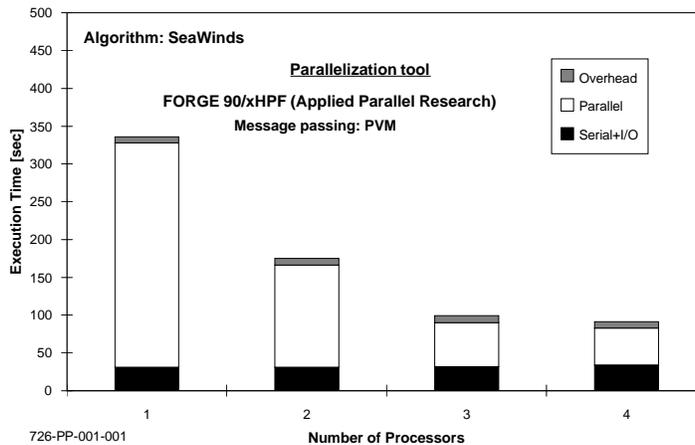
Floating point performance per processor is uniform

NP-7

SeaWinds Execution Time - Workstation Cluster



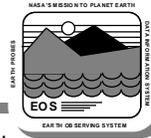
Workstation Cluster (SGI Indigo, HP735, HP735, HP715)
Parallel Workload = 3.2 Giga Floating Point Operations;
Concurrent I/O



Ordering of nodes is important

NP-8

Speedup** - Workstation Cluster

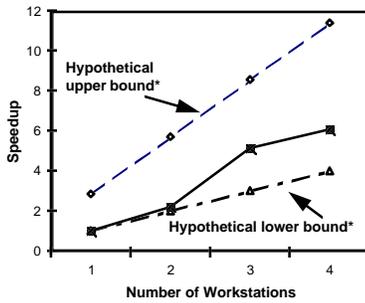


Workstation cluster (SGI Indigo, HP735, HP735, HP715);
Parallel Workload = 3.2 Giga Floating Point Operations;
Concurrent I/O

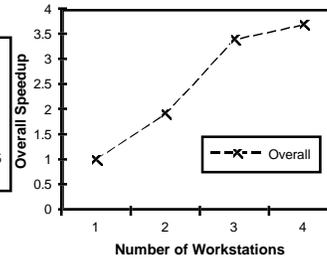
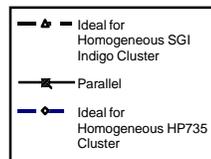
Algorithm: SeaWinds

Parallelization tool

FORGE 90/xHPF (Applied Parallel Research)



Message passing: PVM



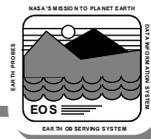
** Speedup is dependent on ordering of nodes normalized to SGI Indigo

* If processors of different clockspeeds are used, we would have a lower bound and an upper bound that encompass the actual speedups. The HP735 was ~3 times faster than SGI Indigo. The slowest machine (SGI) in the cluster determines the lower bound, while the fastest machine (HP735) determines the upper bound.

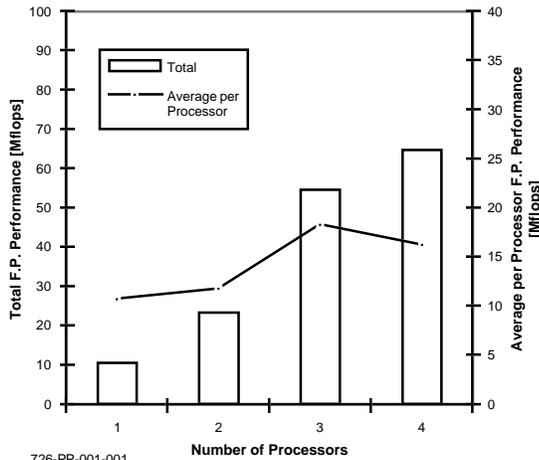
726-PP-001-001

NP-9

Floating Point Performance - Workstation Cluster



Workstation cluster (SGI Indigo, HP735, HP735, HP715); Parallel workload = 3.2 Giga Floating Point Operations; Concurrent I/O



726-PP-001-001

Rated Peak MFLOPS

SGI Indigo : 16
HP735 : 40
HP715 : 13

34% average processor efficiency for 4 processors

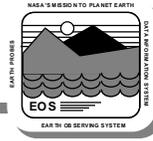
Parallelization tool

FORGE 90/xHPF (Applied Parallel Research)

Message passing: PVM

NP-10

Cray T3D Hardware Platform Specifications

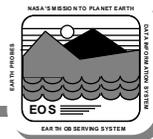


- **Integrated host system**
 - 1 - 4 Cray Y-MP CPUs
 - 333 MFLOPs per CPU peak performance
 - 1 - 3 I/O clusters
- **Cray T3D system**
 - 4.8 - 307.2 GFLOPs peak performance
 - MIMD architecture with SIMD support
- **Processing Elements (PEs)**
 - DEC Chip 21064 64-bit RISC microprocessor
 - 150 MFLOPs per PE
 - 16 or 64 MB local memory per PE
 - 6.6 ns clock speed

726-PP-001-001

NP-11

Cray T3D Hardware Platform Specifications (cont.)



- **Memory**
 - 60 ns, 4-Mbit or 16-Mbit DRAM
 - Physically distributed, globally addressable
 - 2 to 16 GB (single cabinet)
- **I/O**
 - 2 to 4 I/O gateways
 - 1.6 GB/s peak I/O bandwidth
- **Interconnect network**
 - 76.8 GB/s bisectional bandwidth in all directions

726-PP-001-001

NP-12

Cray T3D Development Environment

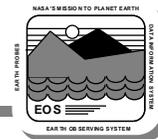


- **CRAFT (Cray Research Adaptive Fortran) programming model supports with worksharing directives**
 - Data parallel programming method (simultaneously operates on many elements of a data structure, typically Fortran arrays)
 - Message passing using Cray optimized PVM (allows programmers the most explicit control over communication among processing elements)
- **CrayTools**
 - advanced programming utilities to help programmers write, analyze, and debug codes for maximum performance
- **MPP Apprentice**
 - high-level performance analyzer
- **TotalView**
 - symbolic source-level debugger for parallel programs
- **Other parallelization tools (e.g. Forge 90/xHPF)**

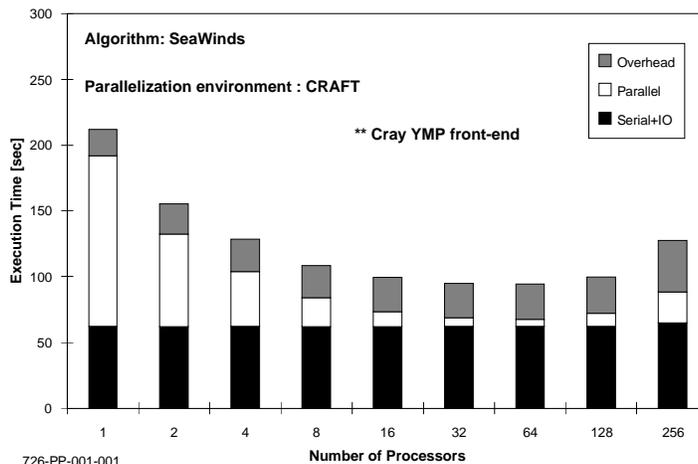
726-PP-001-001

NP-13

SeaWinds Execution Time - Cray T3D**



Execution Time - Cray T3D;
Parallel Workload = 3.2 Giga Floating Point Operations



Good parallel processor

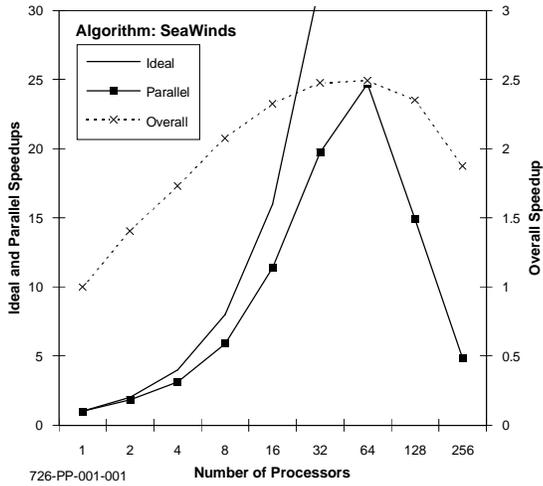
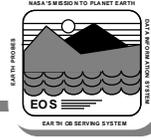
I/O is slow because of front-end Cray YMP

726-PP-001-001

NP-14

Speedup on Cray T3D**

Parallel Workload = 3.2 Giga Floating Point Ops.



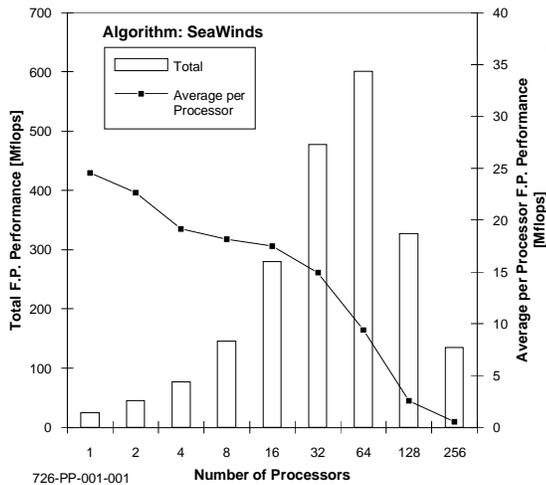
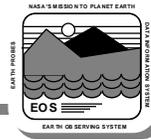
Parallelization environment : CRAFT

** Cray YMP front-end

NP-15

Floating Point Performance - Cray T3D

Parallel Workload = 3.2 Giga Floating Point Operations



Rated Peak MFLOPS : 150 per processing element

Parallelization environment : CRAFT

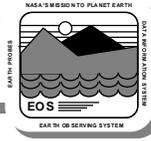
Actual MFLOPS is:

17% of rated peak for 1 processor

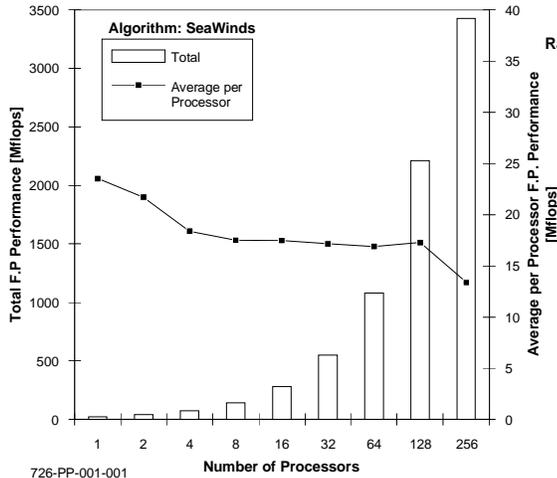
7% of rated peak for 64 processors

NP-16

Floating Point Performance - Cray T3D (cont.)



Parallel Workload = 407 Giga Floating Point Operations



Rated Peak MFLOPS : 150 per processing element

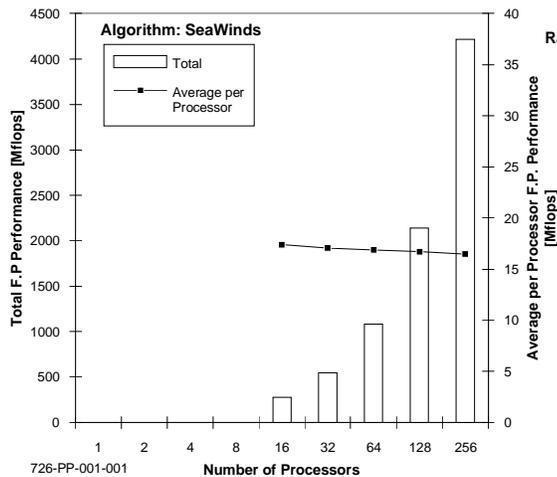
Parallelization environment : CRAFT

NP-17

Floating Point Performance - Cray T3D (cont.)



Parallel Workload = 3261 Giga Floating Point Operations



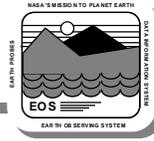
Rated Peak MFLOPS : 150 per processing element

Parallelization environment : CRAFT

Actual MFLOPS approaches 15 MFLOPS/processor (11% of rated peak) for large number of processors)

NP-18

IBM SP2 Hardware Platform Specifications

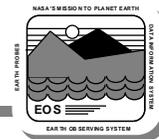


- **SP2 system**
 - number of processor nodes (2 to 128)
 - memory (128 MB to 1/4 TB)
 - Peak performance (0.5 to 34 GFLOPs)
- **Two types of nodes for flexibility**
 - wide nodes have 64 MB to 2048 MB with 66.7 MHz clockspeed
 - thin nodes (half the physical size of the wide nodes) have 64 MB to 512 MB with 66.7/62.5 MHz clockspeed
- **High performance switch is the interconnection network that links all the nodes together**
 - point to point transfer is 40 MBps
 - latency (500 ns to 80 nodes; 875 ns over 80 nodes)
- **Every node can be used for external network communications and I/O connections**

726-PP-001-001

NP-19

IBM SP2 Development Environment

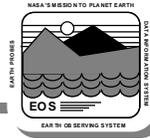


- **Distributed memory processor**
- **Message passing libraries supported**
 - PVM, Express, Linda, MPI, MPL (IBM proprietary)
- **Parallelization tools**
 - currently no native tools
 - supports third-party automated tools like Forge 90/xHPF
- **Parallel debugger**

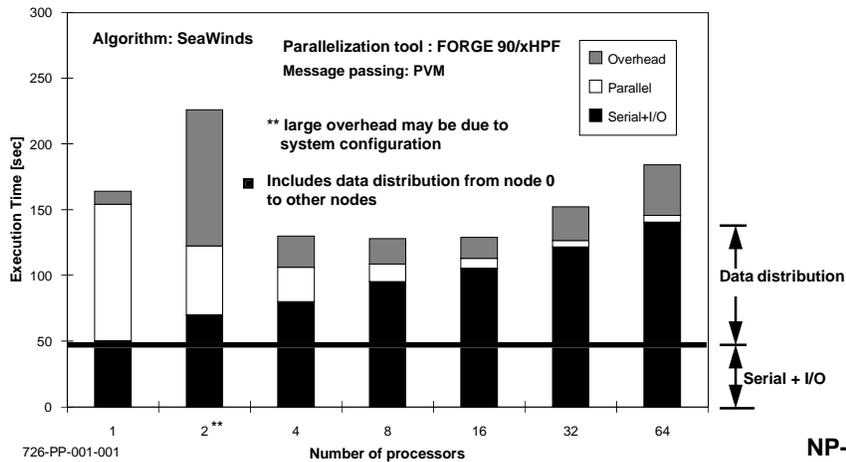
726-PP-001-001

NP-20

SeaWinds Execution Time - IBM SP2



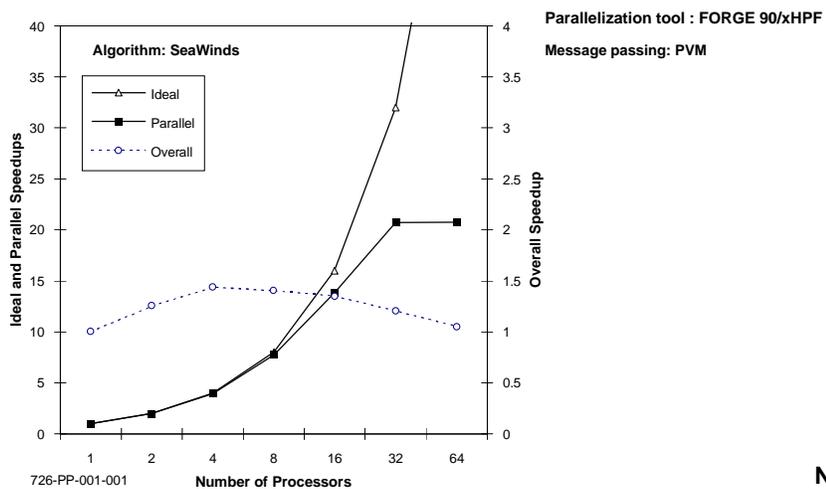
Execution Time - IBM SP2;
Parallel Workload = 3.2 Giga Floating Point Operations



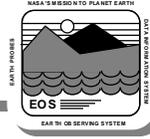
Speedup on IBM SP2



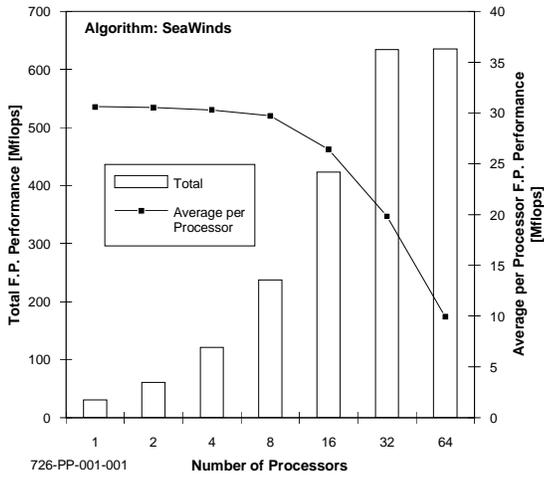
Parallel Workload = 3.2 Giga Floating Point Ops.



Floating Point Performance - IBM SP2



Parallel Workload = 3.2 Giga Floating Point Operations



Rated Peak MFLOPS : 266 per processor

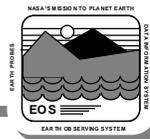
Parallelization tool : FORGE 90/xHPF

Message passing: PVM

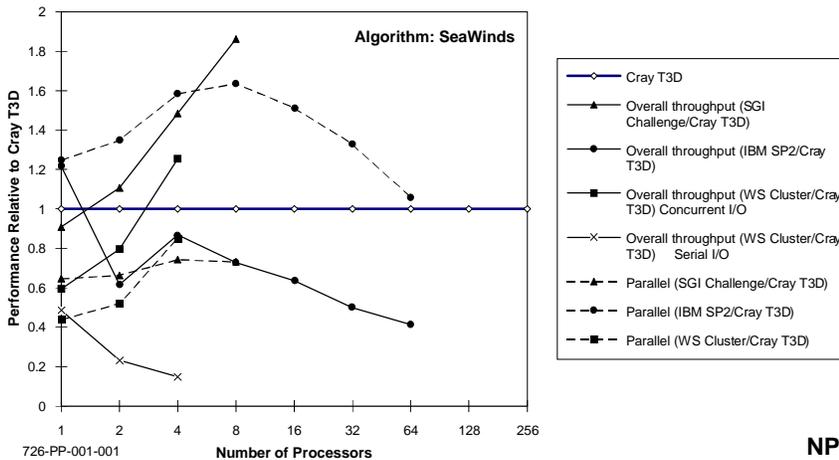
11% of rated peak (1 processor)
4% of rated peak (64 processors)

NP-23

Performance Relative to Cray T3D

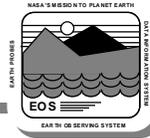


Performance of SGI Challenge, IBM SP2, and Workstation Cluster Normalized to Cray T3D

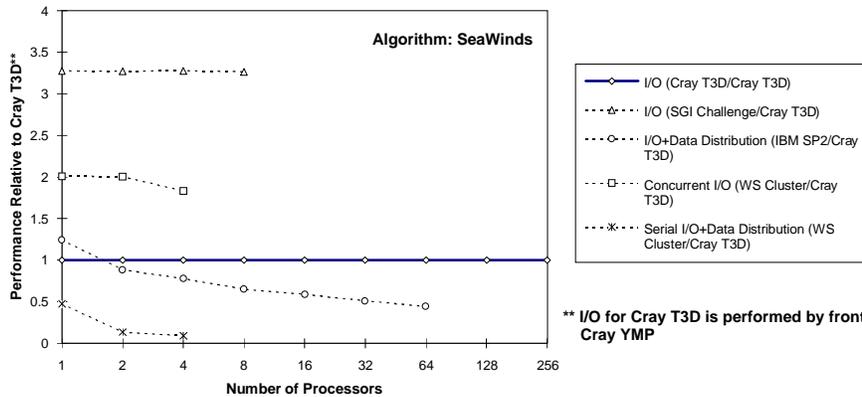


NP-24

I/O and Data Distribution Performance Relative to Cray T3D with Cray YMP front-end



I/O and Data Distribution Performance of SGI Challenge, IBM SP2, and Workstation Cluster Normalized to Cray T3D**

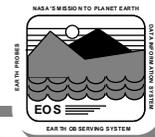


** I/O for Cray T3D is performed by front-end Cray YMP

726-PP-001-001

NP-25

Conclusions

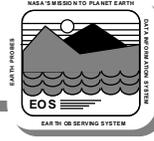


- Parallelization developed for SMP/Workstation cluster is applicable for MPPs
 - SMP/workstation cluster can be used as a starting point
- If same parallelization tool is used, migration to MPP is simplified
- Actual MFLOPS performance for SGI Challenge XL was uniform as a function of processors
- Need very large workloads for executing parallel programs efficiently on MPPs
- Overall throughput (includes I/O, data distribution and numerical operations)
 - SGI Challenge was better than Cray T3D, IBM SP2 or workstation cluster (both using PVM)
 - workstation cluster with serial I/O gave worst performance (speed of interconnect can affect performance)
 - workstation cluster with concurrent I/O appears promising and has the potential to equal the Cray T3D or IBM SP2
- Parallel numerical operations
 - IBM SP2 (using PVM) outperformed Cray T3D for small number of processors, but performance degrades (after attaining a peak) for large number of processors
- I/O performance
 - Front-end machines and disk configuration can impact I/O performance of MPPs
- Data distribution among multiple processors
 - SMP performs better because data distribution is implicit
 - Cray T3D performance is better than IBM SP2

726-PP-001-001

NP-26

Appendix

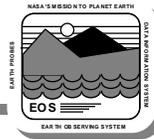


The following slides will not be covered in this presentation.

726-PP-001-001

NP-27

A-1 Purdue Benchmarks - SGI Challenge XL



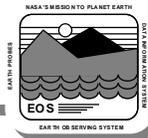
Reference

PDPS Prototyping at ECS Science and Technology Laboratory: Progress Report #4 (Document # 194-00569TPW)

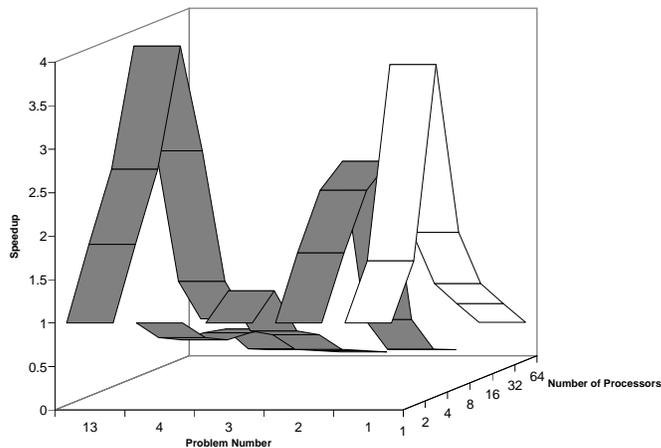
726-PP-001-001

NP-28

A-3 Purdue Benchmarks - Speedup on Cray T3D



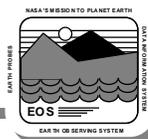
Speedup Ratios for Purdue benchmarks on Cray T3D



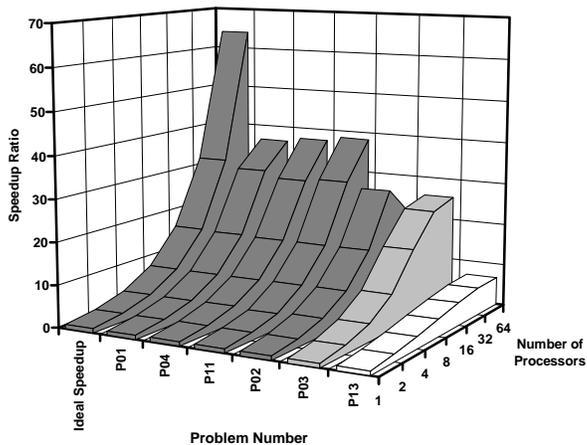
726-PP-001-001

NP-31

A-4 Purdue Benchmarks - Speedup on IBM SP2



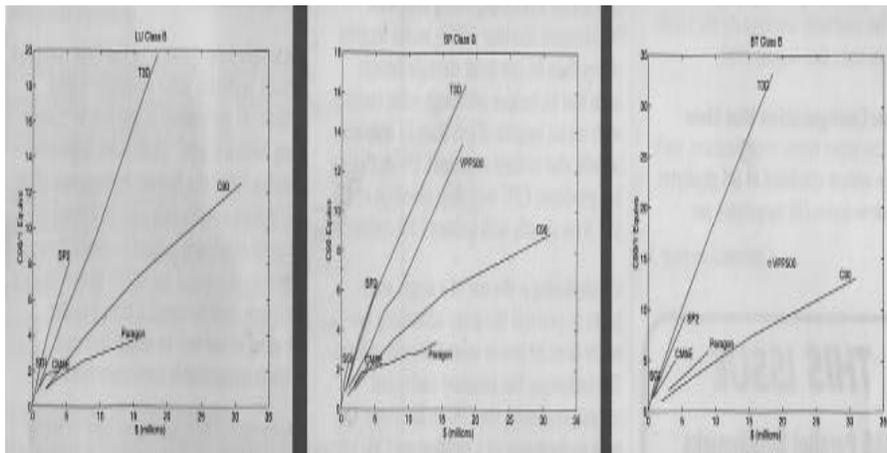
Speedup Ratios for Purdue Benchmarks on NAS IBM SP/2



726-PP-001-001

NP-32

A-5 NAS Results based on Computational Fluid Dynamics Benchmarks**

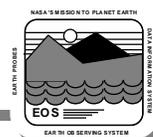


726-PP-001-001

** Reference: NAS News Volume 2, Number 8, January-February 1995

NP-33

A-6 Prototyping Documents



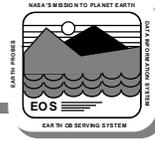
The ECS Data Handling System (EDHS) provides information about all the prototypes. The URL is <http://edhs1.gsfc.nasa.gov/>

- Click SDPS under Segment Office Home Pages
- Click SDPS Prototypes
- This prototype is listed under Science Software Execution Prototype

726-PP-001-001

NP-34

A-7 Acknowledgements



The Cray Supercomputer used in this investigation was provided by Jet Propulsion Laboratory, Pasadena, CA under funding from the NASA offices of Mission to Planet Earth, Aeronautics and Space Science.

The IBM SP2 used in this investigation was provided by NASA/Ames, Moffet Field, CA under funding from the NASA offices of Mission to Planet Earth, Aeronautics and Space Science.

Drs. Steve Gunter and Scott Dunbar of JPL are acknowledged for providing SeaWinds software for ECS prototyping.