NAS Experience with the Cray X1

Rupak Biswas
Subhash Saini, Sharad Gavali, Henry Jin, Dennis Jespersen,
M. Jahed Djomehri, Nateri Madavan, Cetin Kiris

NASA Advanced Supercomputing (NAS) Division
NASA Ames Research Center, Moffett Field, California

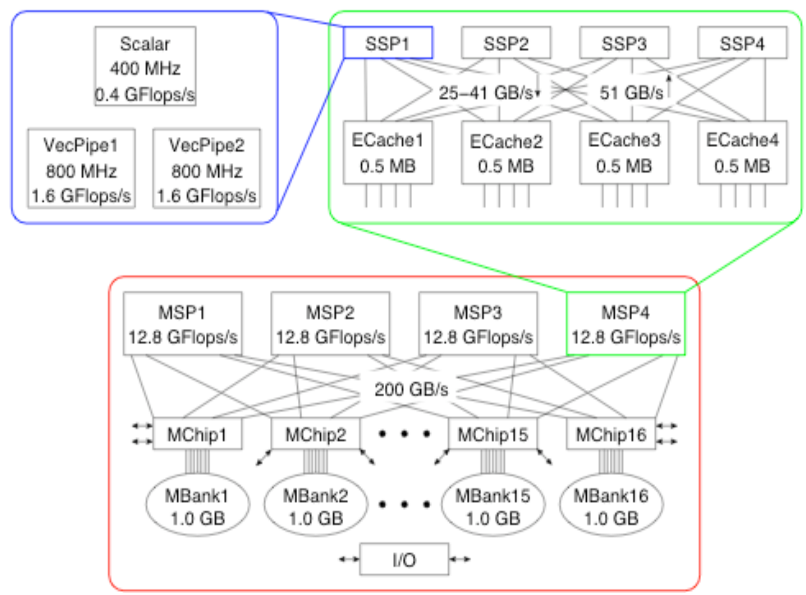
CUG 2005, Albuquerque, May 19

Outline

- **Cray X1 at NAS**
- **Benchmarks**
  - HPC Challenge Benchmarks
  - NAS Parallel Benchmarks
  - Co-Array Fortran SP Benchmark
- **Applications**
  - OVERFLOW
  - ROTOR
  - INS3D
  - GCEM3D
- **Summary**

Cray X1 at NAS

- **Architecture**
  - 4 nodes, 16 MSPs (64 SSPs)
  - 1 node reserved for system;
    3 nodes usable for user codes
  - 1 MSP: 4 SSPs at 800 MHz, 2 MB ECache
    12.8 Gflops/s peak
  - 64 GB main memory;
    4 TB FC RAID



- **Operating Environment**
  - Unicos MP 2.4.3.4
  - Cray Fortran and C 5.2
  - PBSPro job scheduler

Objectives

- **Evaluate spectrum of HEC architectures to determine their suitability for NASA applications**
  - Compare relative performance by using micro-benchmarks, kernel benchmarks, and compact and full-scale applications
  - Determine effective code porting and performance optimization techniques
- **Use suite of testbed systems as gateways to larger configurations at other organizations**
  - NAS recognized expert in single-system image systems
  - Trade Columbia cycles with other supercomputers based on optimal application-to-architecture matching

Evaluation Environment

- **Cray X1**
  - Both MSP and SSP modes
  - MPI, OpenMP, hybrid MPI+OpenMP, Multi-Level Parallelism (MLP), and Co-Array Fortran (CAF) programming paradigms
  - Profiling tools (e.g. pat_hwpc)
- **Compared with SGI Altix (Columbia node)**
  - Itanium2 processor, 1.5 GHz, 6MB L3 cache
  - 512 processors
  - MPI, OpenMP, MLP programming paradigms
  - Intel Fortran compiler

HPC Challenge Benchmarks

- **Basically consists of 7 benchmarks**
  - **HPL:** floating-point execution rate for solving a linear system of equations
  - **DGEMM:** floating-point execution rate of double precision real matrix-matrix multiplication
  - **STREAM:** sustainable memory bandwidth
  - **PTRANS:** transfer rate for large data arrays from memory (total network communications capacity)
  - **RandomAccess:** rate of random memory integer updates (GUPS)
  - **FFTE:** floating-point execution rate of double-precision complex 1D discrete FFT
  - **Latency/Bandwidth:** ping-pong, random & natural ring

HPCC Performance

- Baseline run on 48 processors without tuning or optimization

Benchmark	Units	SGI Altix	Cray X1
G-PTRANS	GB/s	0.890	0.025
G-Random Access	GU/s	0.0017	0.00062
EP-Stream Triad	GB/s	2.488	62.565
G-FFTE	GFlop/s	0.632	0.192
EP-DGEMM	GFlop/s	5.446	9.889
Random Ring Bandwidth	GB/s	0.746	2.411
Random Ring Latency	us	4.555	13.719

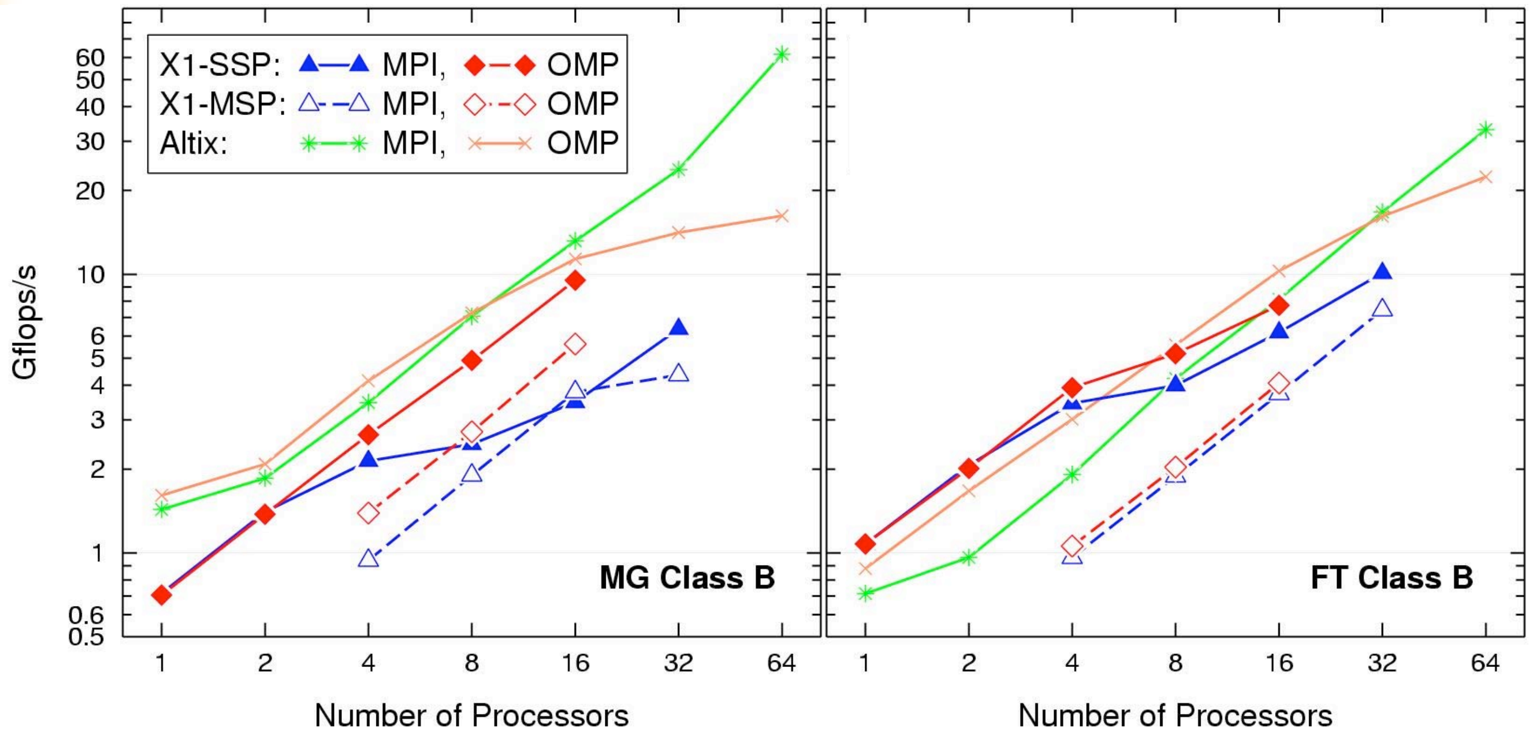
NAS Parallel Benchmarks (NPB)

- Derived from Computational Fluid Dynamics (CFD) applications
- Widely used for testing parallel computer performance
- Five kernels and three simulated CFD applications
- Implemented with MPI, OpenMP, and other paradigms
- Recent work
  - Unstructured Adaptive (UA) benchmark
  - Multi-Zone versions (NPB-MZ)
  - Larger problem sizes

NPBs used in Evaluation

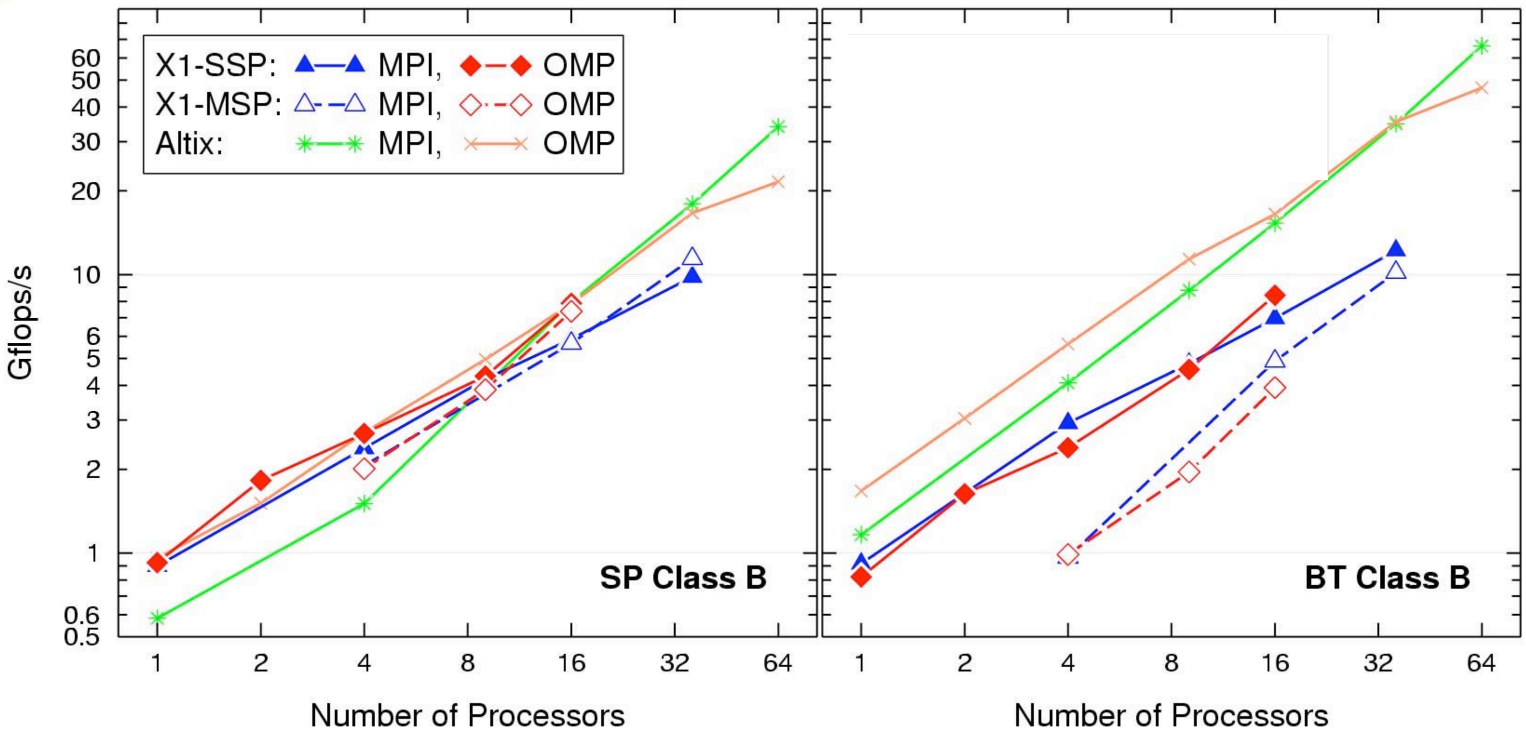
- **Kernel benchmarks**
  - **MG:** multi-grid on a sequence of meshes
  - **FT:** discrete 3D FFTs
- **Application benchmarks**
  - **BT:** block tridiagonal solver
  - **SP:** scalar pentadiagonal
  - **LU:** lower-upper Gauss Seidel

NPB: MG, FT Performance



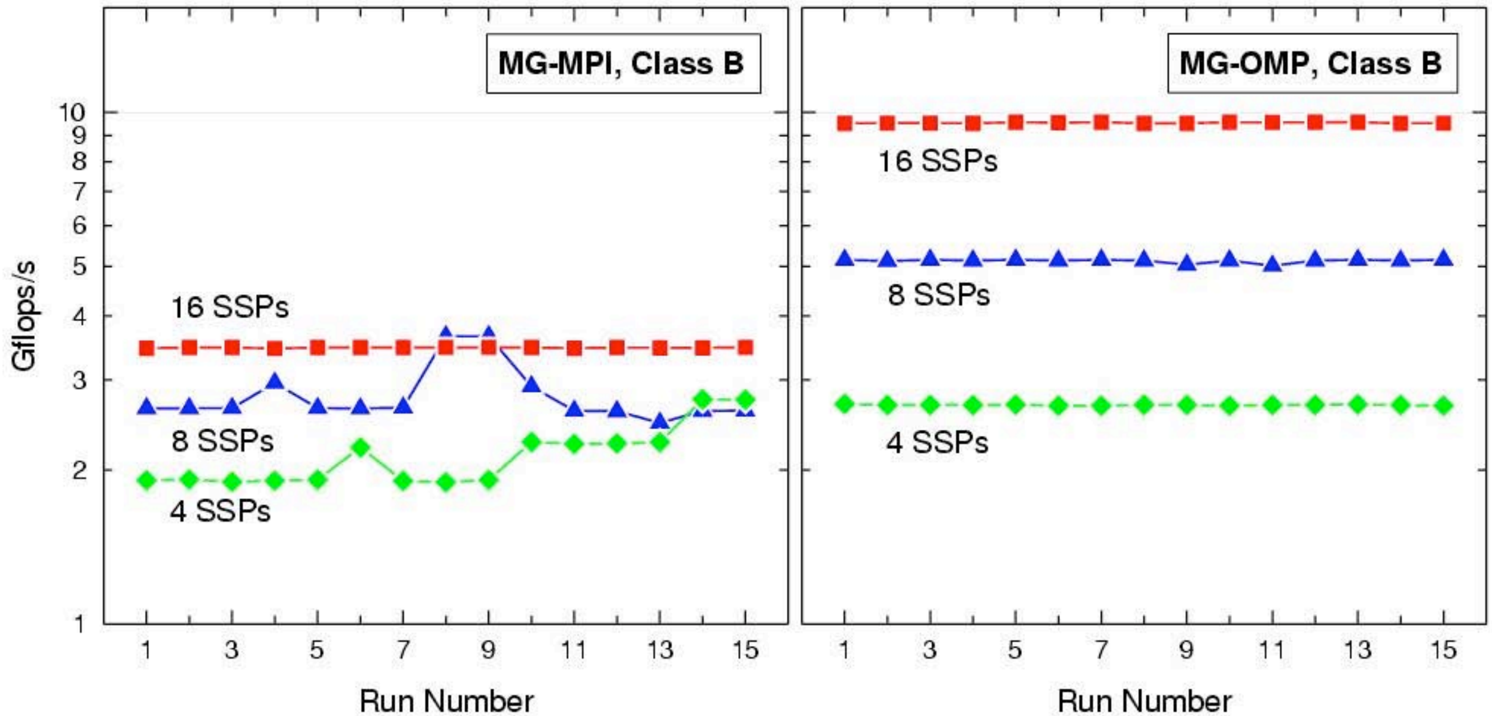
- MPI SSP runs have scaling problem; MPI MSP runs scaled well but showed poor performance
- Streaming is problematic in both benchmarks
- OpenMP shows better performance than MPI on the X1, but reverse is true on the Altix

NPB: SP, BT Performance



- For SP, MPI and OpenMP versions show similar performance in both SSP and MSP modes, indicating proper streaming
- For BT, MSP runs scaled better than SSP runs, but poor streaming
- One Altix processor is equivalent to one X1-SSP for SP, but Altix doubled performance for BT

NPB: Timing Variation in SSP Runs



- Large timing variation in MPI SSP runs when number of SSPs not a multiple of 16
- No similar problem observed in OpenMP SSP runs

NPB: Single Processor Performance

- Reported by pat_hwpc

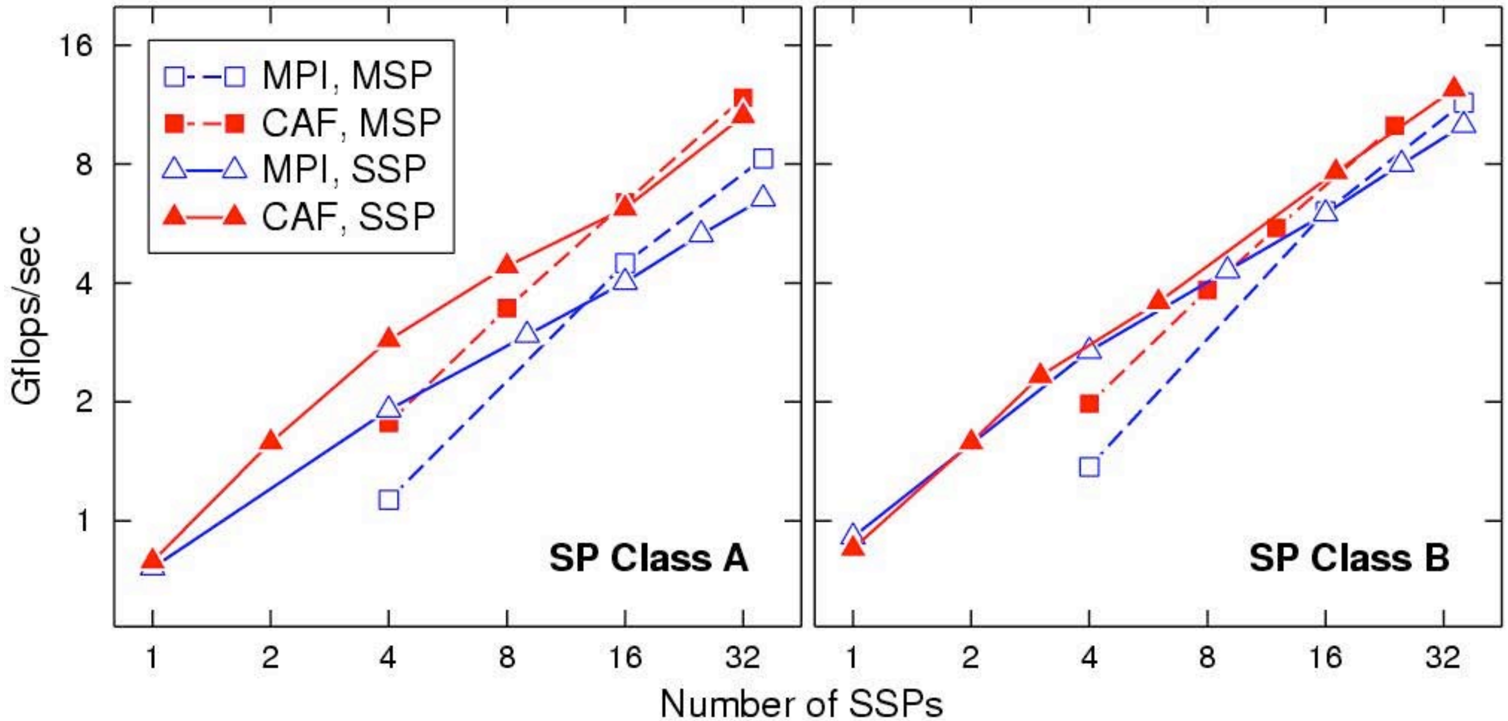
NPB	FP/ Load	Avg. Vec. Len.		% of Peak	
		SSP	MSP	SSP	MSP
MG.B	0.85	44.65	27.21	24.0	11.8
FT.B	0.94	64.00	17.49	35.5	8.3
SP.B	1.10	49.88	35.37	28.9	17.4
BT.B	0.95	60.87	49.76	25.8	8.3
LU.B	1.75	55.71	42.80	35.5	21.8

- Poor “floating-point operations per load” numbers directly impact performance, especially in MSP mode
- Reduced average vector length in MSP runs indicate streaming affects vectorization, causing MSP performance degradation

Co-Array Fortran (CAF) SP Benchmark

- CAF is a robust, efficient parallel language extension to Fortran95
- Shared-memory programming model based on one-sided communication strongly recommended for X1
- Evaluate CAF by creating a parallel version of the SP benchmark from NPB 2.3
- Start from scratch: serial vector version
- Run class A and class B problem sizes
- Compare results with MPI vector version

CAF SP Performance



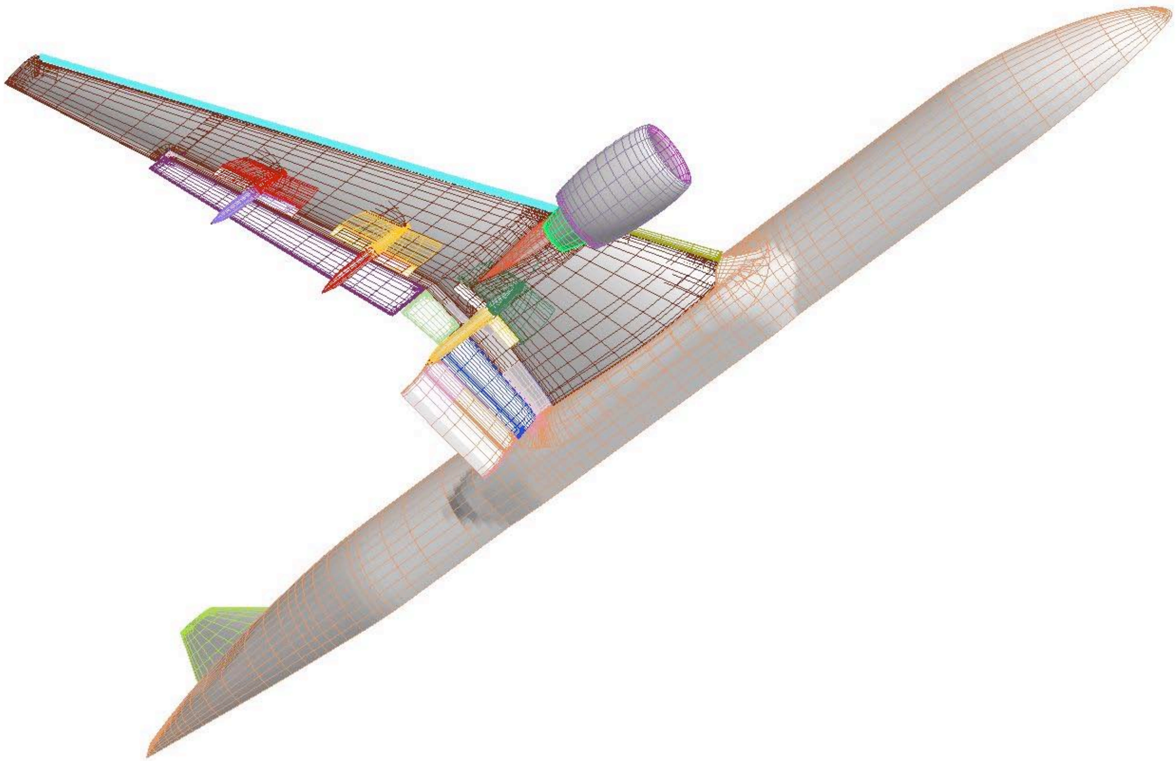
- CAF shows consistently better performance than MPI
- In SSP mode, CAF version also scales better
- MSP runs show worse performance on small processor counts, but outperform SSP runs for large numbers of processors

Application: OVERFLOW

- NASA's production CFD code
- Fortran77, ~100,000 lines, ~1000 subroutines
- Development began in 1990 at NASA Ames
- Solves Navier-Stokes equations of fluid flow with finite differences in space and implicit integration in time
- Multiple zones with arbitrary overlaps (boundary data transfer using Chimera scheme)
- Cray vector heritage
- Multi-Level Parallelism (MLP) paradigm
  - Forked processes using shared memory for coarse-grain parallelism across grid zones (blocks)
  - Explicit OpenMP directives for fine-grain parallelism within grid zones

OVERFLOW Test Case

- Realistic aircraft geometry: 77 zones, 23 million grid points



OVERFLOW Performance

- Average wall clock per step, hardware performance monitor

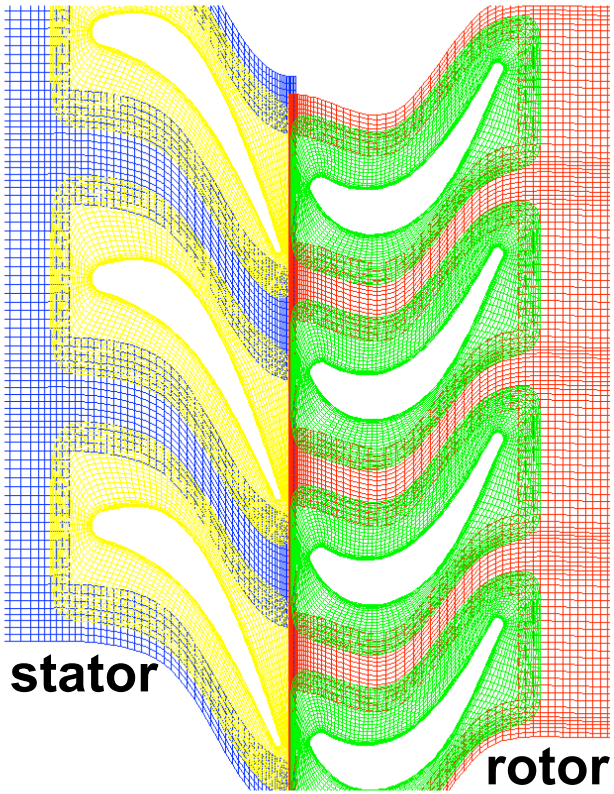
SGI Altix		Cray X1				
CPU	sec/step	MSP	sec/step	Gflops/s	FP/Load	Avg. Vec. Len.
4	19.871	1	26.205	2.895	1.39	50.20
8	9.893	2	13.215	5.462	1.39	50.11
16	5.235	4	6.869	9.763	1.39	49.80
32	2.784	8	3.481	15.885	1.38	49.23
48	2.152	12	2.343	19.666	1.38	48.25

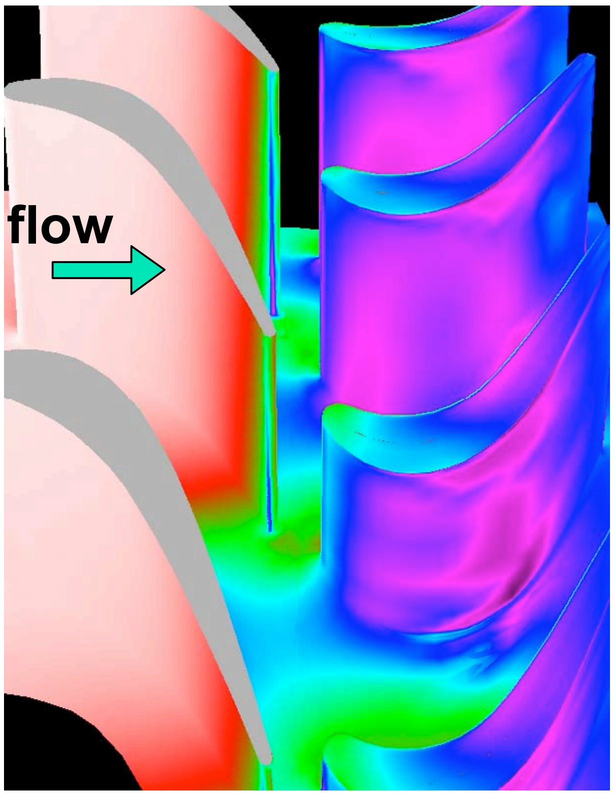
- All X1 runs in MSP mode; OpenMP replaced by streaming; a few explicit directives were necessary
- One MSP roughly equivalent to 3.5 Altix CPUs
- Reasonable vector length, but low FP operations per memory load
- 23% of peak on one MSP; 20 Gflops/s and 13% of peak on 12 MSPs
- Better scaling on X1 than Altix

Application: ROTOR

- Multi-block, structured-grid CFD solver for unsteady flows in gas turbines
- Developed at NASA Ames in late 1980, early 1990
- Basis of several unsteady turbomachinery codes in use in government and industry
- Solves Navier-Stokes equations in time-accurate fashion
- Uses 3D system of patched and overlaid grids and accounts for relative motion between grids

ROTOR Test Case





multiple zone grid system

3D grid formed by stacking multiple 2D grids & wrapping around cylindrical surface

pressure distribution

unsteady flow due to relative motion; rotor interaction with wakes & vortices; vortex shedding from blunt trailing edges

ROTOR: “Serial” Performance on X1

- Serial code optimized for C-90; 6 airfoils, 12 grids, compiler-generated automatic streaming

Case	Size	Mode	Time (s)	FP/ Load	Vec. Len.	Gflops /s	% of Peak
Coarse	0.7M	MSP	12.10	1.15	20.1	0.77	7.8
		SSP	16.94	1.24	24.9	0.53	16.6
Medium	6.9M	MSP	79.42	1.17	30.2	1.32	10.3
		SSP	135.30	1.26	38.1	0.75	23.4
Fine	23.3M	MSP	225.62	1.18	36.6	1.58	12.3
		SSP	414.24	1.26	42.1	0.83	25.9

- Serial code runs more efficiently in SSP mode
- 8-12% of peak performance achieved in MSP runs; 17-26% for SSP
- Code vectorizes well; but average vector lengths higher in SSP mode

ROTOR: MSP vs. SSP Performance

- Both MLP and CAF versions with 12 processors and one OpenMP thread

Case	Size	Mode	Para-digm	Time (s)	Gflops /s	% of Peak
Coarse	0.7M	MSP	MLP	1.00	7.30	4.75
		SSP	MLP	2.03	3.60	9.38
		MSP	CAF	0.99	7.38	4.80
		SSP	CAF	1.90	3.85	10.03
Fine	23.3M	MSP	MLP	20.61	15.37	10.00
		SSP	MLP	49.40	6.41	16.69
		MSP	CAF	20.21	16.34	10.64
		SSP	CAF	47.66	6.65	17.32

- Both CAF and MLP implementations run more efficiently in SSP mode
- CAF version shows about 5% better performance than MLP
- Performance improves with bigger problem size

ROTOR: MLP vs. CAF Performance

- Effect of multiple OpenMP threads (in SSP mode)

Case	Size	SSP	OMP Thrd	MLP+OpenMP			CAF+OpenMP		
				Time (s)	GF/s	SU	Time (s)	GF/s	SU
Coarse	0.7M	12	1	2.03	3.60	1.00	1.90	3.85	1.00
		24	2	1.06	6.90	1.92	1.02	7.15	1.86
		36	3	0.72	10.10	2.81	0.70	10.51	2.73
		48	4	0.57	12.80	3.56	0.55	13.32	3.45
Fine	23.3M	12	1	49.40	6.41	1.00	47.66	6.65	1.00
		24	2	23.76	13.33	2.08	22.97	13.79	2.07
		36	3	17.24	18.38	2.87	16.49	19.21	2.89
		48	4	13.88	22.82	3.56	12.77	24.81	3.73

- Efficiency of about 90% going from 12 to 48 SSPs
- CAF still better than MLP (both with multiple OpenMP threads)
- More OpenMP threads or MSP mode not evaluated due to machine size

ROTOR: X1 vs. Altix Performance

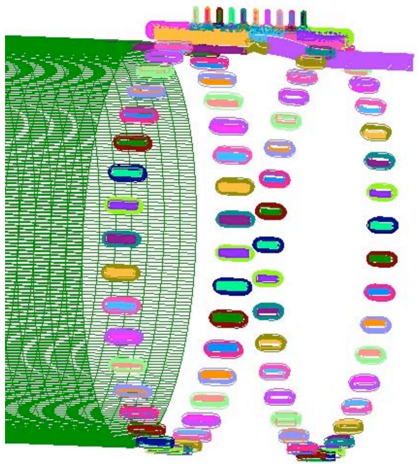
- CAF+OpenMP in SSP mode on X1; cache-opt MLP+OpenMP on Altix

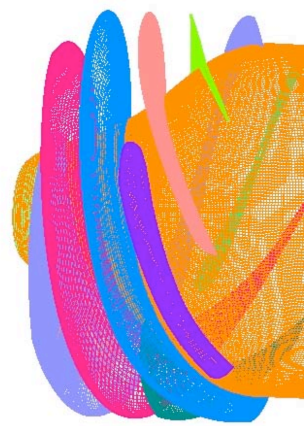
Case	Size	CPU	OMP Thrd	Cray X1			SGI Altix		
				Time (s)	GF/s	SU	Time (s)	GF/s	SU
Coarse	0.7M	12	1	1.90	3.85	1.00	1.69	4.32	1.00
		24	2	1.02	7.15	1.86	1.04	7.07	1.64
		36	3	0.70	10.51	2.73	0.95	7.74	1.79
		48	4	0.55	13.32	3.45	0.84	8.68	2.01
Medium	6.9M	12	1	17.01	5.62	1.00	19.29	4.95	1.00
		24	2	8.23	11.61	2.07	11.08	8.62	1.74
		36	3	5.99	15.96	2.84	10.12	9.44	1.91
		48	4	4.60	20.78	3.70	8.89	10.75	2.17

- OpenMP scales better on X1 than Altix (little speedup beyond 2 threads)
- For small problem sizes that fit in cache, Altix has slight advantage; however, X1 outperforms for larger problems

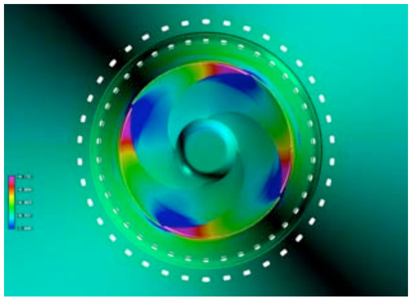
Application: INS3D

- High-fidelity CFD for incompressible fluids
- Multiple zones with arbitrary overlaps (overset grids)
- Cray vector heritage
- Hybrid programming paradigm
  - MPI for coarse-grain inter-zone parallelism
  - OpenMP directives for fine-grain loop-level parallelism
- Flow Liner Analysis
  - 264 zones, 66M grid points
  - Smaller case for X1:
    only S-pipe A1 test section
    (6 zones, 2M grid points)



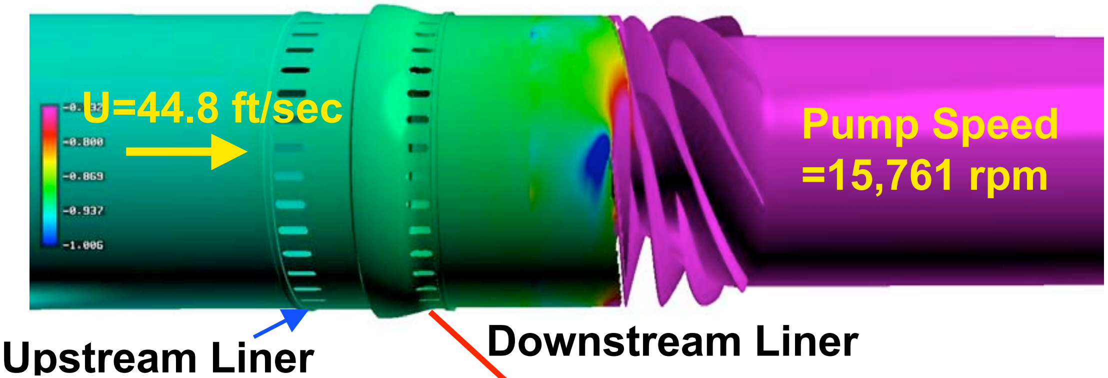


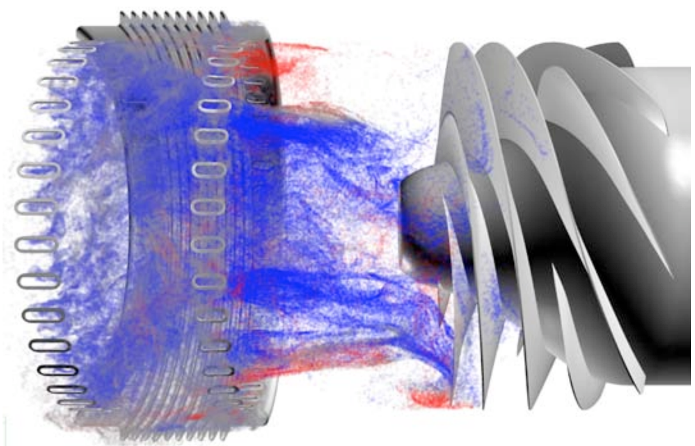
INS3D Test Case



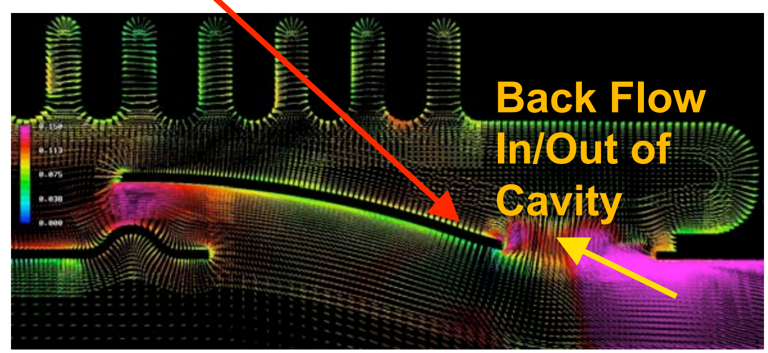
Strong backflow causes HF pressure oscillations

Unsteady Simulation of SSME LH2 Flowliner



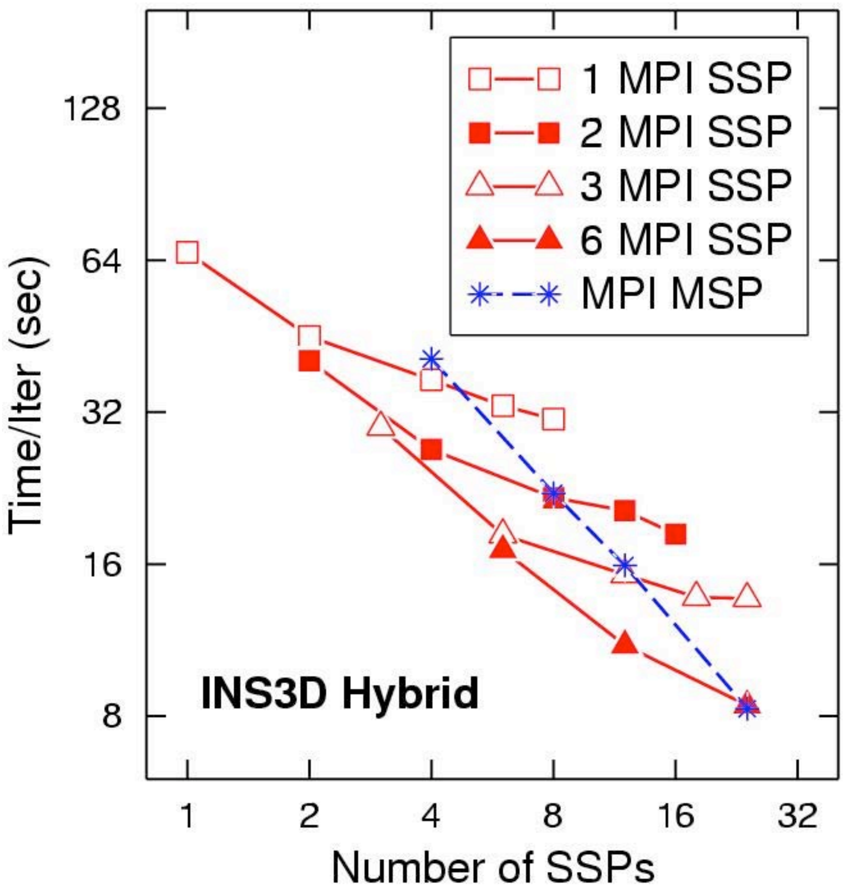


Particle traces colored by axial velocity values (red indicates backflow)



Damaging frequency on flowliner due to LH2 pump backflow has been quantified in developing flight rationale

INS3D Performance



- 6 zones grouped into 1, 2, 3, 6 MPI groups
- For each group parameter value, used 1, 2, 4, 6, 8 OpenMP threads in SSP mode
- MPI scaled well in SSP mode up to the 6 groups
- OpenMP scaling deteriorated after 4 threads
- Performance in MSP mode similar to SSP case using 4 OpenMP threads, indicating streaming in MSP was as effective as OpenMP

INS3D Performance

	SSP mode; 1 MPI			SSP mode; 4 OMP			MSP mode; MPI		
SSP	**Time (s)**	**Vec. Len.**	**FP/ Load**	**Time (s)**	**Vec. Len.**	**FP/ Load**	**Time (s)**	**Vec. Len.**	**FP/ Load**
1	66.9	42.7	1.90						
2	45.2	36.6	1.87						
4	37.1	28.6	1.78	37.1	28.6	1.78	40.8	19.2	1.91
8	31.0	20.7	1.64	21.7	28.9	1.85	22.1	19.2	1.94
12				15.3	28.6	1.84	15.9	19.2	1.94
24				8.4	28.6	1.88	8.3	19.2	1.95
	0.994 — 2.148 Gflops/s			**1.792 — 7.968 Gflops/s**			**1.623 — 8.265 Gflops/s**		

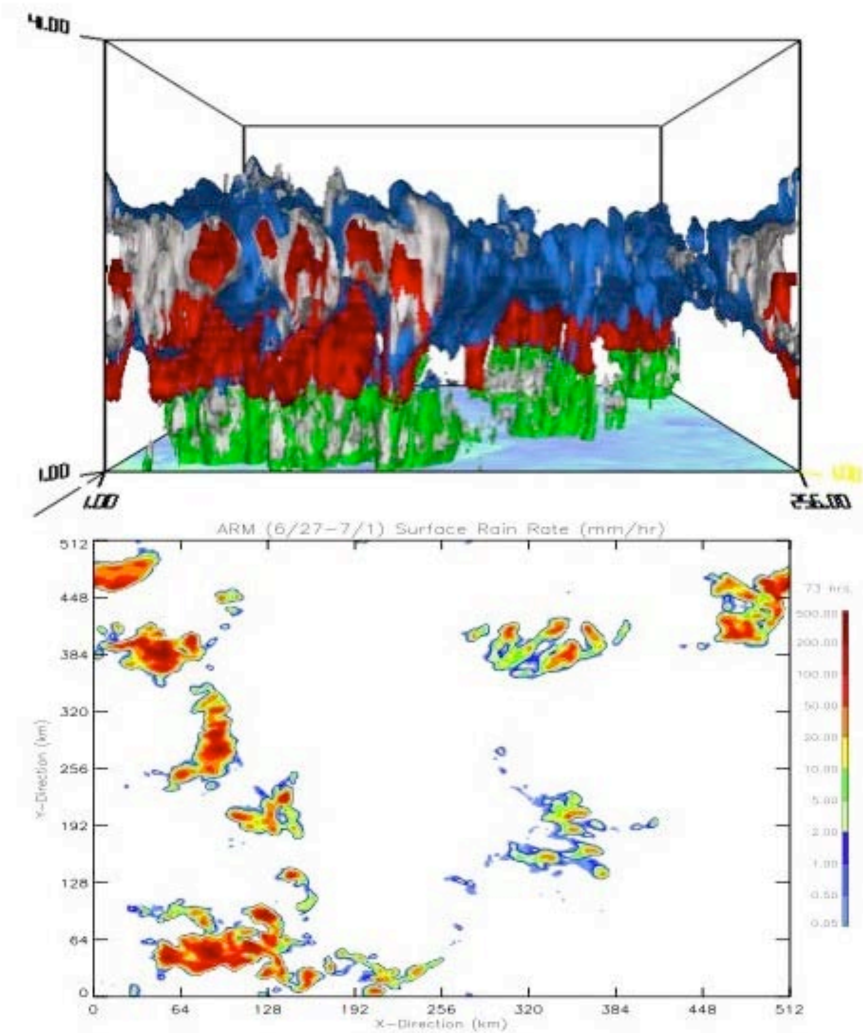
- Good vector operations per memory load (~1.9)
- 31% of peak on 1 SSP; 14.0%–9.8% in SSP mode with 4 OpenMP threads; 12.7%–10.8% in pure MPI mode on MSP

Application: GCEM3D

- Goddard Cumulus Ensemble (GCE) model
  - Regional cloud resolving model, developed at GSFC
- Two parallel versions of GCEM3D
  - MPI code is coarse-grain parallelization based on domain decomposition strategy
  - OpenMP code is fine-grain parallelization at loop level
- Both versions solve the same geophysical fluid dynamics models, except that the land component is included only in the MPI version at this time
- MPI domain decomposition in 2-D; longitude and latitude directions
- Optimization achieved by compiler flags and directives

GCEM3D Test Case

- Linear cloud system propagating from west to east in SCSMEX (S. China Sea)
  (by Tao et. al.)
- Cloud isosurfaces
  - White: cloud water and cloud ice
  - Blue: snow
  - Green: rain water
  - Red: Hail
- Surface rainfall rate (mm/hr)



GCEM3D Performance

	MPI / MSP (104x104x42)			MPI / SSP (104x104x42)			OpenMP / SSP (256x256x32)		
SSP	Time (s)	Vec. Len.	FP/ Load	Time (s)	Vec. Len.	FP/ Load	Time (s)	Vec. Len.	FP/ Load
1				2519	47	2.1	5872	56	1.7
2				1401	46	2.1	3102	55	1.7
4	1772	27	2.0	920	45	2.0	1647	53	1.7
8	901	27	2.0	523	45	2.0	880	50	1.7
12				524	24	2.0			
16	675	15	1.9	413	24	2.0	475	44	1.7
32	342	15	1.9	237	24	1.9			

- MPI code in MSP/SSP modes scales well up to 8 MSPs/SSPs
- MPI in SSP mode about 2x better than MSP (better vector length in absence of multistreaming)
- OpenMP scales better, but worse sustained performance (lower operations per load, lower vectorization ratio, missing land model)

Summary

- Relatively user-friendly programming environment, effective compilers and tools, several programming models available
- Two different modes, MSP and SSP, provide additional flexibility in parallelization and tuning
- For the test suite, 25% of peak easily achieved in SSP mode, but automatic streaming in MSP mode not as effective
- Co-Array Fortran straightforward to implement and offered improved performance over MPI and MLP
- OpenMP scaling reasonably well up to four threads
- Timing variations observed in SSP mode believed to be related to the X1 design
- Preliminary comparison between X1 and Altix indicate equivalent performance between SSP and Itanium2