# Challenges in Application Scaling In an Exascale Environment

14<sup>th</sup> Workshop on the Use of High Performance Computing In Meteorology

November 2, 2010

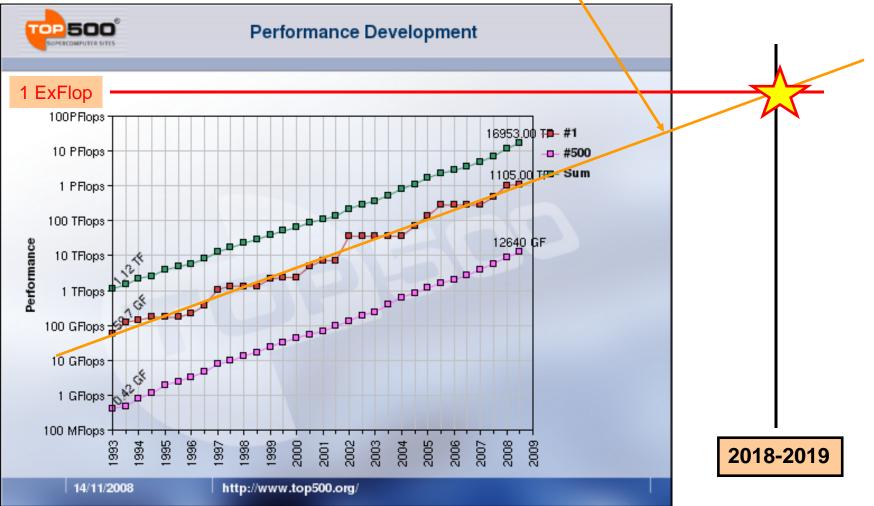
Dr Don Grice IBM

Page 2



### Growth of the Largest Computers by year 1000x every 10-11 years

### ExaFlop in 2018-2019!





# Technology Trends

# Different Scaling Trends for Different Technologies Compute Ratios will Change

- Driven by Cost and Energy Usage
- Circuit-Flop Densities will Continue to Improve
- I/O BWs and Power will not improve as quickly
  - Technology Improvements may help this
  - Costs may still be limiters
- Memory Volume Costs (and BWs) may be Limiting

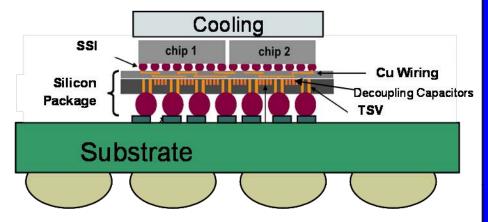


## The Big Leap from Petaflops to Exaflops

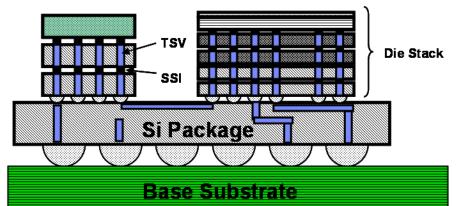
- Technology disruptions in many areas driven by Power and Cost Concerns.
- All Impact System Balance and Application Optimization
  - Silicon power scaling:
    - Frequency Plateau more threads needed
    - Impacts Application Scaling, Power Usage, and RAS
  - Memory technology Volume and BW
    - Bytes/Flop ratios decreasing (Locality Counts)
  - Interconnect BW
    - Bytes/Flop ratios decreasing (Locality Counts)
  - Packaging technology I/O Switching Costs
    - Relative Amount of Power needed to move Data Increasing
- Need to be able to exploit machines. Not just about flops.
   Flop metric promises to be an even poorer predictor of sustained performance than it is now

# 3D Si Pkg & Si Die Stacking

#### **3D Silicon Pkg Integration**



#### 3D Die Stack & Si Pkg Integration



#### 3D Microsystems Technology:

- 1. Through-silicon-vias (TSV) & Thin Si
- 2. Fine pitch wiring / interconnection
- 3. Integrated active devices / passive function
- 4. Silicon-Silicon Interconnection (SSI) die stacks & Si pkg (C2C, C2W, W2W)
- 5. Test Known good die & Die stacks
- 6. Power delivery / advanced cooling

#### <u>Opportunities:</u>

- High bandwidth
- Low Power
- Heterogeneous chips
- Modular Design
- Modular System Integration
- System Power Savings
- System Performance Gain
- System Miniaturization
- Time to Market
- Lower cost



## DARPA IPTO Sponsored (Information Processing Techniques Office) ExaScale Software Study: Software Challenges in Extreme Scale Systems September 14, 2009

"This document reflects the thoughts of a group of highly talented individuals from universities, industry, and government research labs on the software challenges that will need to be addressed for the Extreme Scale systems that are anticipated in the 2015 – 2020 time-frame."

Used as Reference Point for some Application Challenges



## **Locality Tension/Trade-offs**

- Energy to move data will be significant
- Application run time optimization may run counter to keeping things local
  - Convergence time reduction and Locality may be at odds
  - AMR may disrupt locality
  - Multi-Scale and Multi-Physics may disrupt neighbor locality
  - Unstructured Grids disrupt Neighbor Communication Patterns
  - Global Data Access may aid convergence
- Models need to consider:
  - Asynchronous lightweight tasks and communications
    - Latency Hiding
  - Explicit Locality Model
  - Scalable Coordination and Synchronization
  - Abstract Performance Model: Parallelism, Locality, Energy



# Reliability

- New technologies can be leveraged to deliver a reliable system even at an Exascale.
  - The biggest challenge is that architectures must be willing to pay the price of hardware detection.
  - There will be a tension between energy efficiency and error detection
  - Do we need a software construct like a 'critical section' that must get the 'right answer'?



# **Petascale/Exascale Computing**

# **Software Challenges**

# Types of Application Scaling

- Strong Scaling
  - Making the same problem run in less time or running more time steps in the same time.
  - Not many applications exhibit strong scaling
- Weak Scaling
  - Increasing the problem size to run on a larger machine in a 'tractable time'. (e.g. Spatial Scaling)
  - Finer meshes generally imply smaller time steps
  - Creates an issue when a problem has a fixed window to run in
- 'New Era' Scaling
  - Increasing the amount of science computed during the solution of the problem.



# "New Era" Weak Scaling Not Just Solving Things Faster – Solving Them Better

"New-era" weak scaling typically adds extra work through one or more of the following:

- Multi-scale
- Multi-physics (multi-models)
- New models
- Interactions
- Mitigation analysis
- Data mining
- Data-derived models

# **Scaling Limitations**

Not all applications will scale to exascale with their current structure due to things like:

- Parallelism
  - O (10\*\*11) Threads required
  - Load Imbalance and Serial Portions
- Locality
  - Vertical (temporal)
    - Data Reuse is not always possible
    - Movement in the Memory Hierarchy Occurs
  - Horizontal (data decomposition)
    - Excessive Movement uses Energy
    - Introduces Latency and Serialization Issues
- Bottlenecks in Memory, Communications, and I/O BWs

#### IBM

## Exascale Software Challenges: Summary

- Dealing with Frequency Stabilization
  - Increased number of threads for additional performance
  - Load Balancing is critical for Sustained Performance Gains
- Dealing with Performance Ratio Changes
  - Computation will be 'free'
  - Recomputing vs moving data may be the best solution
- Dealing with Memory Footprint issues
  - A view of exascale machines as a memory system with embedded compute facilities may be helpful
- Debugging and Optimizing Codes at Scale
  - New Tools and methods are evolving



# HPC Software Tools What's being done to Help with Scaling?

### IBM

### **10X System Wide Productivity Improvements**

#### Programmer

- Eclipse Tools
- Compiler Enhancements
- UPC Language
- Automated Performance Tuning

### **Administrator**

- Automated Discovery
- Automated Configuration
- Diskless Boot
- Rolling Updates
- Server, Network, & Storage

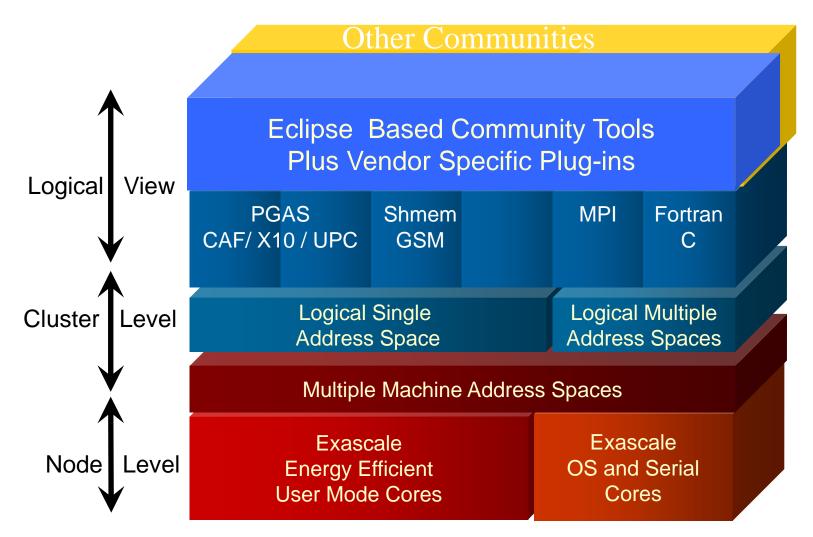
### **System Operational Efficiency**

- Resource & Workload Management
- Protocol Optimizations & PGAS
- Co-scheduling
- Dynamic Page Size Assignments

### **Reliability and Serviceability**

- Continuous Operation
- Checkpoint/Restart
- Server, Network, & Storage
   Monitoring
- Problem Determination

## **Application Writer's View of the System**



# IBM PERCS Goals

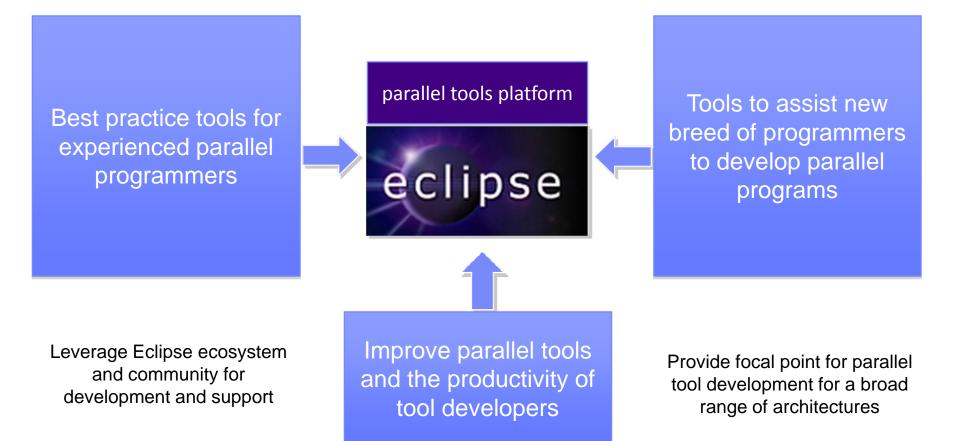
## Productive, Easy to Use, Reliable Computing System

- Collaborative ecosystem based on Eclipse
  - Parallel Tools Platform
  - Integrated tools
- Multi-language support
   C, C++, UPC, Fortran, X10
- Platform independence
- Highly scalable
  - 1M Tasks
- Enhanced productivity
  - -Lower entry point for new users



# Parallel Tools Platform (PTP)

### Enabling Parallel Application Development



## Contributors

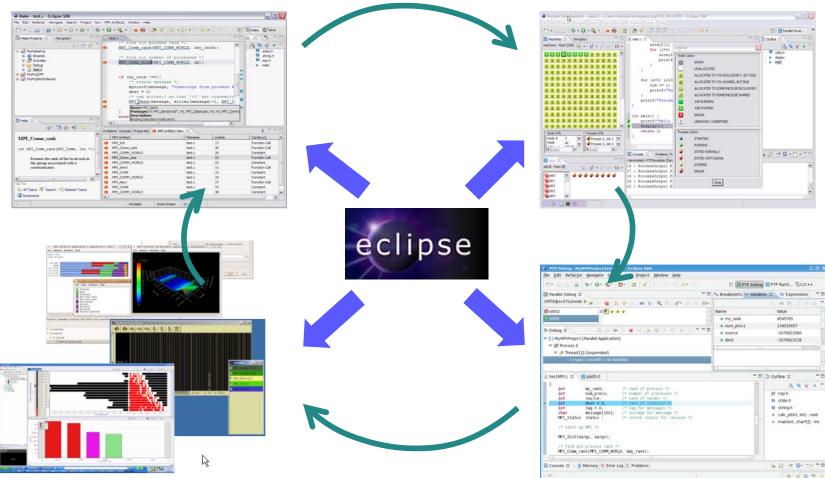
- IBM
- Los Alamos National Laboratory
- Oak Ridge National Laboratory
- University of Tennessee at Knoxville
- University of Oregon
- Monash University
- National University of Defense Technology
- NCSA
- University of Illinois
- University of Utah
- Munich University of Technology
- Forschungszentrum Juelich

IBM Systems and Technology Group

# **HPC** Application Development Cycle

#### **Coding & Static Analysis**

### **Application Execution**



#### **Dynamic & Performance Analysis**

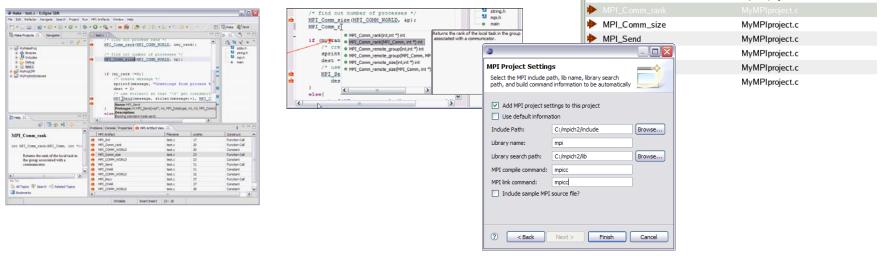
### **Application Debugging**

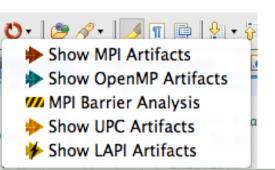
© 2010 IBM Corporation

## Coding & Static Analysis

## Parallel Language Development Tools (PLDT)

- Assistance tools to increase productivity of parallel programmers
  - New project wizards (MPI, OpenMP)
  - Content Assist (command/API completion), hover help, builtin API help descriptions in an html help "view" (MPI, OpenMP, LAPI, UPC)
  - Location of parallel "artifacts" in code: MPI, OpenMP, LAPI APIs, and UPC





Filename

MyMPIproject.c

MyMPIproject.c

MyMPIproject.c

🖉 Tasks 📮 Console 🚼 Problems 🐤 MPI Artifact View 🖾

Artifact

MPI\_Bcast

MPI\_Init

MPI\_Reduce

Page 23

## **Application Execution**

### Launching & Monitoring

PTP Runtime - job02:1 - Eclipse 5DK     File Edit Navigate Search Bun Project Window Help		Create, manage, and run configurations Create a configuration to launch a par	allel application in Parallel Perspective
	and instance of the local state		
C+ S S S S S S O+ Q+   E ] /   H+ S+ S O+ O+ O	E PTP Runti C/C++	I ₪ × ⊨ ∌•	Name: shallow
Resource Managers 22 = C	etestMPI.c Mob02:1 E	type filter text ③	Resources Application & Debugger 🚧 Arguments 😓 Source 🖉 Environment 🗔 Common
🙀 ORTE@sc07tutorial (ORTE)	Process details	C/C++ Local Application	Resource Manager: OMPI@LargeCluster
	Index: 1 Node: node1 Total: 4 PID: 21935 Iob: iob02	<ul> <li>Eclipse Application</li> <li>Java Applet</li> </ul>	
	Status: EXITED	Java Application Ju JUnit	Launch Attributes
Machines 22 Q Q Q* (f (c) (b) ~ = 0	- Program output	Jij JUnit Plug-in Test	Simple Advanced
ORTE@sc07tutorial: localhost.localdomain - Root [64]	hi from 1	OSCI Framework     Terratel Application	Number of processes: 1
iocalhost.localdomain 0 🖬 🙆 🙆 🙆 🖉 🚽 🚽	-	iii shallow Jii Task Context Plug-in Test	Options
20		JU Task Context Test	By slot No oversubscribe No local
		Q View Latex Document	Prefix:
50		-	Hosts
60 🔜 🔤 🔤			Host file: Browse
Node Attributes Process Info			Host list:
Attribute Value #job02:1			
Name nodel			
Node Number 1			(Apply) (R
Node State UP		Filter matched 12 of 12 items	
2			Close
🖬 jobs 🕱 🔍 🔍 🔍 👘 🗗 🖉 🖉 👘		Ŷ	LIUSE
ORTE@sc07tutorial: default:job02 - Root [4]	🖸 Consol 👔 Problem 🗖 Properti 🕢 Tasks 🥂 Error Lo 🧊 Progress 🛛 🥂 🗖		
O job02 0 ₽ ♥ ₽ ₽	Rubber understate (fielder)		
	Building workspace (Finished) 36 OK		
	J		
0°	Index: 1 EXITED PID: 21935	(	
		DBS/Tor	
noroves visibilitv i	nto target system	PBS/Tor	que
• •	nto target system	L	que
• •		L	
• •	nto target system	L	
ingle point of inte		L	
ingle point of inte			
ingle point of inter ontrol	rface for launching and	Open M	
ingle point of inter ontrol	rface for launching and	Open M	
ingle point of inter ontrol		Open M	SLURM IPI LoadLeveler
ingle point of inter ontrol anages interactio	rface for launching and	Open M	
ingle point of inter ontrol	rface for launching and	Open M	SLURM IPI LoadLeveler

# Application Debugging PTP Parallel Debugger

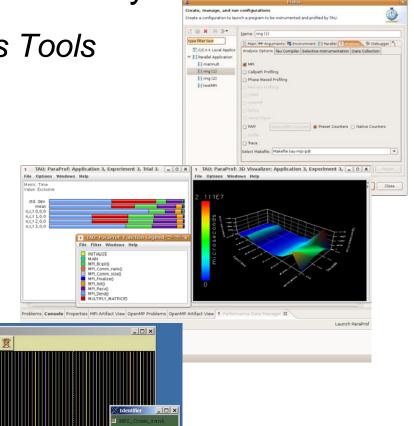
PTP Debug - MyMPIProject/testMPI.c - Eclipse SDK		_ <b>_</b> ×
<u>F</u> ile <u>E</u> dit Refac <u>t</u> or <u>N</u> avigate Se <u>a</u> rch <u>R</u> un <u>P</u> roject <u>W</u> indow <u>H</u> elp		
] Ħ▾ 📓 🚔 📓 ] 券▾ Ο▾ Q▾ ] O་   ☵ ] 🖋 ] 웹< Ϡ< Ⴊ ⇔∙ ⇔<	😰 🗱 PTP Debug	#PTP Runti 🗟 C/C++
🗱 Parallel Debug 🛿 🗖 🗖	In the second secon	🛿 🙀 Expressions 🗖 🗖
ORTE@sc07tutorial: d 📦 🗊 🚰 🧏 🥱 🖃 🔞 0 🔍 🖓 0 👘 🖓 🖉		5) 44 E 5° X % ▽
🤬 job02 0 € → → →	Name	Value
læ job06		4545765
	num_procs	134515657
梦 Debug 22 🦓 🖗 🕪 🖩 🖬 🤱 🥱 论 க i 한 🕱 ་ ▫ ▫	source	-1076822980
マ 詳 MyMPIProject [Parallel Application]	🔷 dest	-1076823128
▽ 🦑 Process 0	( <u> </u>	<b>)</b>
▽ 🖞 Thread [1] (Suspended)		
≡ 1 main() testMPI.c:50 80489b0		
	C	
🖻 testMPI.c 🕱 🏾 🇱 job05:0	- 0 8	🗄 Outline 🕱 🛛 🗖 🗖
<pre>{     fint my_rank;    /* rank of process */     int num_procs;    /* number of processes */     int source;    /* rank of sender */     int dest = 0;    /* rank of receiver */     int tag = 0;    /* tag for messages */     char message[100];    /* storage for message */     MPI_Status status;</pre>		142
<pre>/* start up HPI */ MPI_Init(&amp;argc, &amp;argv); /* find out process rank */ MPI_Comm_rank(MPI_COMM_WORLD, &amp;my_rank); </pre>		
🔄 Console 🕴 🚺 Memory 🔮 Error Log 🛣 Problems		🔒 🛃 🛃 ⊑× 📬 ⊓ 🗖
□ <b>□</b> <sup>•</sup> □		🐴 🤰 🗷 👻 🔶

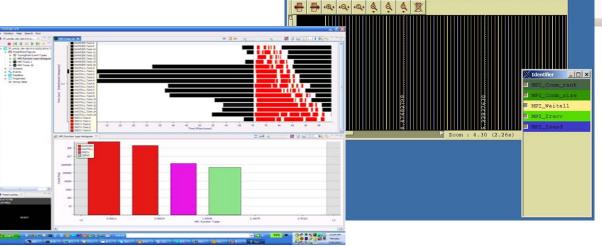
- Mid-scale integrated debugger
- Tightly integrated with Eclipse
- Supports debugging multiple jobs simultaneously
- Utilizes backend debugger (e.g. gdb) for low level operations
- Targeted at SPMD programming models
- Supports mixed MPI & thread debugging
- Single process and group operations
- Platform for building new debugging paradigms

## **Dynamic & Performance Analysis**

## Dynamic Analysis Tools

- Perform analysis on the running application using external tools
- Generate results that must be brought back into Eclipse as part of the development workflow
- May require external tool for visualization or other purposes





# Dynamic & Performance Analysis

### HPC Toolkit

- Integrated with Eclipse and PTP
- Application and tools launched using PE resource manager

the Edit Delactor Hawipole Scarth Project Inst	inumentation Trading Dars NAPI Field Assist Window Help			
: 🖦 - 🔙 🗠 : 🗛 - 👘 : 🛥 🛷	• 10・1回転10・2・0・4・4・			ET 28 femile Spires
Contract State State	· □]]£ desc files (50 - [2] dester as (50 - [2] odd	start 190 (1) meshain 190 (1) mesering 190	(M) develops 100, 32 N, 141 standard laws	1
(post/states) total free deeps/brokepsb		1 1 5 6 1		
	<pre>cnot toc(11, 13, 62)=cnot</pre>	<pre>icc(i1,j3,k2) + cvelgit*isi*igd*isi loc(i2,42,k2) + cvelght*fs2*fv2*fs2</pre>		A
Denterhonene Data (140 2	endl f eps. co			
1.51	Delete T			
See Address in 1,8 1,8 hours deep a Andres Mary Constitution and	aa = (aaa+2) = (aaa+2) = (aaa+2)	total   grad	ipoa a te	
(a) A start of the second s		. NO. PPI_ERAL, NET_SON, C. NET_COM	NOUND, Interi	
And the lower half the lower from the local diapet parallel. A site	1 call My1 REDUCE (show loc, show	e, ng, NFI REAL, MFI SUM, 0, MFI COS		
Jyszyki kysz is 100 (Jelszer (przys / kateria) politie – 4 sk pyszyki kysz is 100 (Jelszer (przys / kateria) (opti prz ile – 5 sk)				6
Jepublickov h (A Uhlwordpepelluteria/hpi profile 6 via Jepublickov h (A Uhlwordpepelluteria/hpi profile 7 via	end suproutine dedsities			-
he data	<u></u>			18.
ne data	- vertamence tata - tree wey 12			-
Line Line V Linewei Lind Line	max dete belon is fer renic 0			
Logend Colleges Select All Dep				Visualize the t
1 G R Inte alls MOD program 8	in land	Number 1	Developed Balan	
E nel el testy 107	C pepele#W	No. of Sec.	In the second se	Performance (
<ul> <li>H-Biggets_sts_H+00_perspen_i+</li> <li>Mins_stanty_2/2</li> </ul>	D proprint proprio 2001			1 chiomanoo .
<ul> <li>[9] np. of kerty_22</li> <li>[1] [9] _ http://dikenty_23</li> </ul>	HPL_Conv_stat_20 L	0.000001	2.000000 🔎	
Page and 242	MALCONV_real_ov 5	0.00057	2.000000	
e nei besst 273	NPL Days 87	5.717557	4723,000000	
- 10 np. (No. 245	NPI Read 2	0.000055	5.00000	
		14,057940	296.060060	
- Muno al tool prov	NEL Altest 240			
[7] mpt_official_2_228.	resused www	0.004020	2522344.000000	
<ul> <li>Very starts 28.</li> <li>Very starts 10.000 photost</li> </ul>		0.004620	2022344.00000	
<ul> <li>Imply that 2,000</li> <li>Imply that 2,000 photon</li> <li>Imply photon<td>resigned wo</td><td>COLUMN T</td><td>A CONTRACTOR OF A CONTRACTOR OF</td><td></td></li></ul>	resigned wo	COLUMN T	A CONTRACTOR OF	
<ul> <li>□ Projectivaliz_200</li> <li>□ Projectivaliz_200</li> <li>□ Projectivaliz_200</li> <li>□ Projectivalizativali</li> <li>□ Projectivalizativali</li> <li>□ Projectivalizativali</li> <li>□ Projectivalizativali</li> <li>□ Projectivalizativali</li> </ul>	resused www	COLUMN T		
<ul> <li>Implication (20)</li> </ul>	resigned wo	COLUMN T	A CONTRACTOR OF	
Comparison of the state o	resigned wo	COLUMN T	A CONTRACTOR OF	
Imp_risks[_00]     model(%)     model(	resigned wo	COLUMN T	A CONTRACTOR OF	Converting
<ul> <li>□ Traje_iterati_con</li> <li>□ Traje_iterati_con</li> <li>□ Traje_iteration</li> <li>□ Traje_itera</li></ul>	resigned wo	COLUMN T	A CONTRACTOR OF	S → We Serve and S → We Serv
Image (and state)_come	resigned wo	COLUMN T	A CONTRACTOR OF	→ V Storm and V S
<ul> <li>[7] republication (2016)</li> <li>[8] republication (2016)</li> </ul>	resigned wo	COLUMN T	A CONTRACTOR OF	■ 2 Max Community of Ant Community of Max Community o
Importantization           Importanization           Importani	resigned wo	COLUMN T	A CONTRACTOR OF	⇒      →
For explosion of the second secon	resigned wo	COLUMN T	A CONTRACTOR OF	Constraints and a second
Imply planting the statig the state of the sta	resigned wo	COLUMN T	A CONTRACTOR OF	King and a second
The spectra with spe	resigned wo	COLUMN T	A CONTRACTOR OF	Kit Conv. (m)     Kit Con
Imply planting the statig the state of the sta	resigned wo	COLUMN T	A CONTRACTOR OF	
For explosion (2006)	resigned wo	COLUMN T	A CONTRACTOR OF	
Imply planting the state growth         Imply planting the state of the stat	resigned wo	COLUMN T	A CONTRACTOR OF	
Projektowie 200	resigned wo	COLUMN T	A CONTRACTOR OF	
Imply planting the statig product of the state of th	resigned wo	COLUMN T	A CONTRACTOR OF	
Importantization           Importanitistion           Importan		COLUMN T	A CONTRACTOR OF	Visualize th
Project kalogene       Pro	resigned wo	COLUMN T	A CONTRACTOR OF	

**P**Gage 26

- Provides an integrated framework for performance analysis
- Looks at all aspects of performance (communication, memory, processor, I/O, etc) from within a single interface
- Operates on the binary and yet provides reports in terms of sourcelevel symbols
- Full source code traceback capability

## Fortran Development Tools

#### Photran project

- http://eclipse.org/photran
- Tech lead at UIUC, Jeff Overbey
- More UIUC students are contributing
- Merged with PTP in 2009
- Photran 6.0 released with PTP 4.0 Helios

#### Photran features:

- Supports Fortran 77-2008
- Syntax-highlighting editor
- GUI interface to gdb
- Makefile-based compilation
- Compiler error extraction
- Outline view
- Open declaration
- Fortran refactorings
- C preprocessor support

Fortran - bstfit.f90 - Eclipse Pl	atform	_ 8 2
<u>F</u> ile <u>E</u> dit <u>N</u> avigate Se <u>a</u> rch <u>P</u> roje	ct <u>R</u> un <u>W</u> indow <u>H</u> elp	
] ◘ + 🔛 👜 📄 ] ≫ + ] 🏇 • 🖸	·• Q.• ] Ē• @• Ē• ] @ ⋪ ] @• ] ⋈ • ⋈ • ↔ ↔	
🖶 Fortran Projects 🛿 📃 🗖	🗈 *bstfit.f90 🕱 🗖	
수 수 🗟 🖻 💲 🔻	module Mdataset	<sup>a</sup> ≷ ≷ ● ▽
D 🔂 MMFP	implicit none	🗧 🗢 🕡 Mdataset 🔺
▽ 📅 > org.eclipse.photran-samp	Peclare a Tdataset type to hold data points	🗢 🕡 Tdataset
Binaries	type :: Tdataset	datapt
▷ 🖂 src-fixed-form	real :: datapt(4,1000)	• npts _
▷ A src-fortran-and-c	integer :: npts real :: valsum(4)	valsum
▷ A src-gaussian-elimination	real     :: valsum(4)       real     :: valsum(4)	valsum
▷ Src-includes	real :: xysum(3) logical :: has np(2)	xysum
Src-language-features	end type Tdataset	has np
	and model a Markensk	
Image: Section of the section of	end module Mdataset	LoadDataSet
.cvsignore 1.1 (ASCI	program BstFitProj	ShowDataPoints
♦ 🚑 src-modules 💌	Suse.Mdataset	SquareDistances
O Halle Hallgets D	Problems      Console Properties	
h (+ - + ) 📐	26 errors, 0 warnings, 0 infos	
▷ ≌org.eclipse.photran-samples	Description	Resource
▷ 😂 whatever2	😣 Derived type definition has no components	bug138221.f
	😣 Error: Component 'valsum' at (1) already declared at (2	2) bstfit.f90
	8 FILE tag must be of type CHARACTER	bug118974.f
	😣 Illegal preprocessor directive	bug118961.f 🗸
		Þ
□		Ć.



## Fortran Refactoring

efactor	τûR
Rename	τôΜ
Extract Procedure	TOL
Extract Local Variable	
EXTRACT LOCAL	107
Introduce IMPLICIT NONE	
Encapsulate Variable	
Encapsulate	
Make Private Entity Public Add ONLY Clause to USE Statement	
in imize UNL. Cans	istent
Add ONLY List Minimize ONLY List Make COMMON Variable Names Const Move SAVE Variables to COMMON Blog Move SAVE Variables to COMMON Blog	LK
Move SAVE Variables	
Move SAVE Variables Interchange Loops (Unchecked)	TOC
Interchange	
Unify Keyword Case Replace Obsolete Operators	P
Replace	
(Debugging)	
(Der	

### Some samples:

- Rename
  - Change name of every use of a variable, function, etc.
  - Only proceeds if new name will be legal
- Extract procedure
  - Moves statements into a new subroutine, replacing statements with a call
  - Local variables are passed as arguments
- Introduce implicit none
  - Adds an 'implicit none' statement
  - Adds explicit variable declarations
- Currently 16 refactorings

Image courtesy Jeff Overbey, UIUC

## Conclusion

- Fundamental Programming Style not likely to change much
  - Multi-Threaded 'MPI tasks' will be the norm
  - New Languages are emerging to help with extreme scale
- A Shared Memory model at the Task level will still exist
- Amount of Threading will have to increase
- 'More Science' will be a way to use cycles
- Optimization Points will change Computing is 'free'
- New Tools are emerging to help create applications 'at scale'

IBM Systems and Technology Group

## In the end, it's not about the technology; It's what you do with it that counts

# Thank you...

# ...any Questions?

© 2010 IBM Corporation