Next Generation Performance Tools: A Framework for Automated Performance Bottleneck Detection

David Klepacki
Advanced Computing Technology
IBM T.J. Watson Research Center

Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.

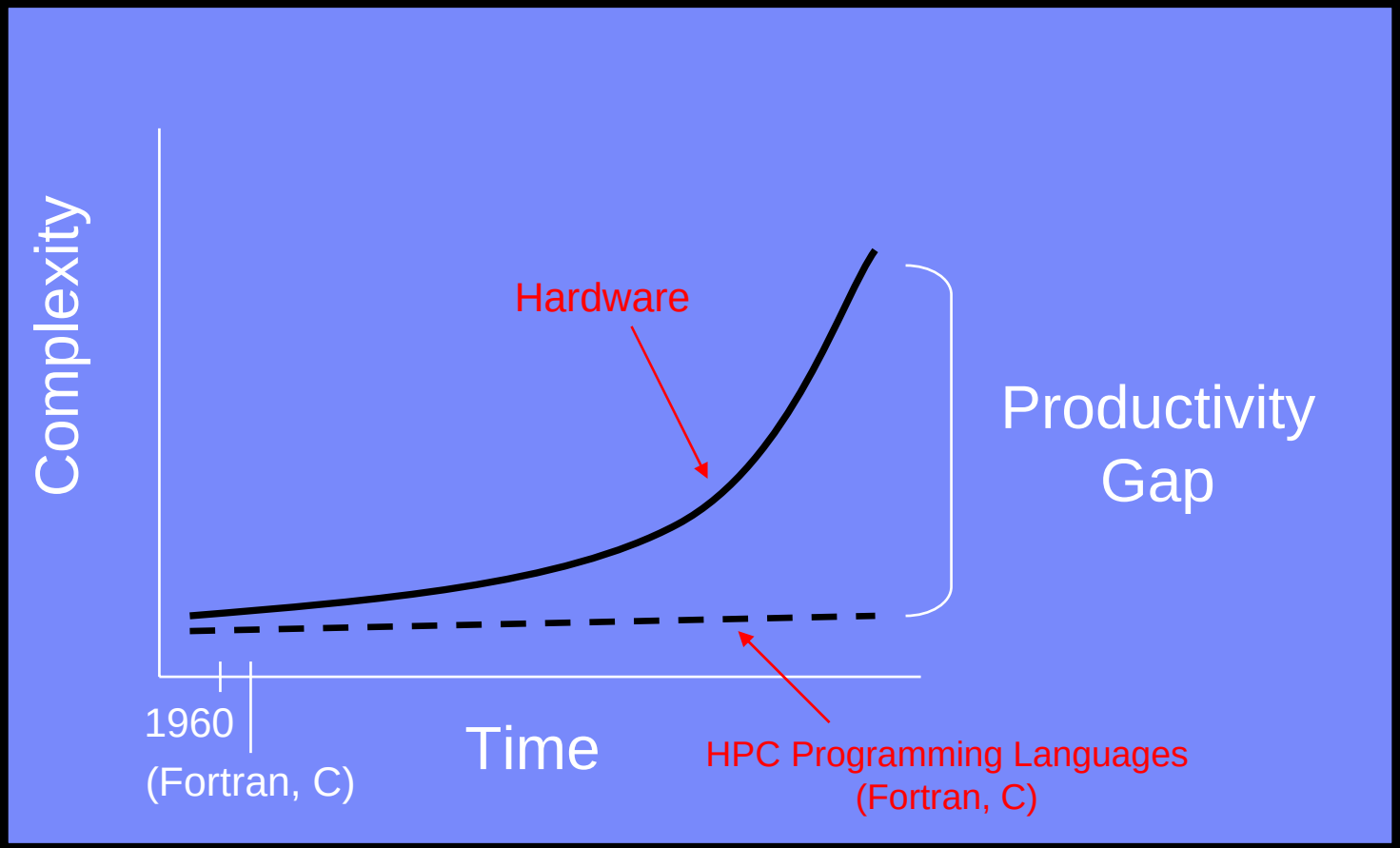
© 2007 IBM Corporation

System Evolution

- **Device Scaling imposing fundamental constraints on system**
  - Power dissipation and energy consumption
  - Physical size / packaging
- **Pressure to re-think system architecture**
  - Blue Gene: low power devices, embedded (small)
  - Cell: Attached (embedded) co-processing engine
- **Systems become inherently more complex**
  - Connectivity / hierarchical topology (torus, intra-cell)
  - Multi-core processors (and less memory per processor)
  - Multi-thread (SMT, hyperthreading)
- **This poses new challenge to application programming**
  - New programming paradigm? (but ~$1T in legacy codes, ISV apps, etc.)
- **Conclusion: New software tools essential to mitigate evolving system complexity and improve productivity.**

Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

Enablement Productivity Gap = Hardware – Software



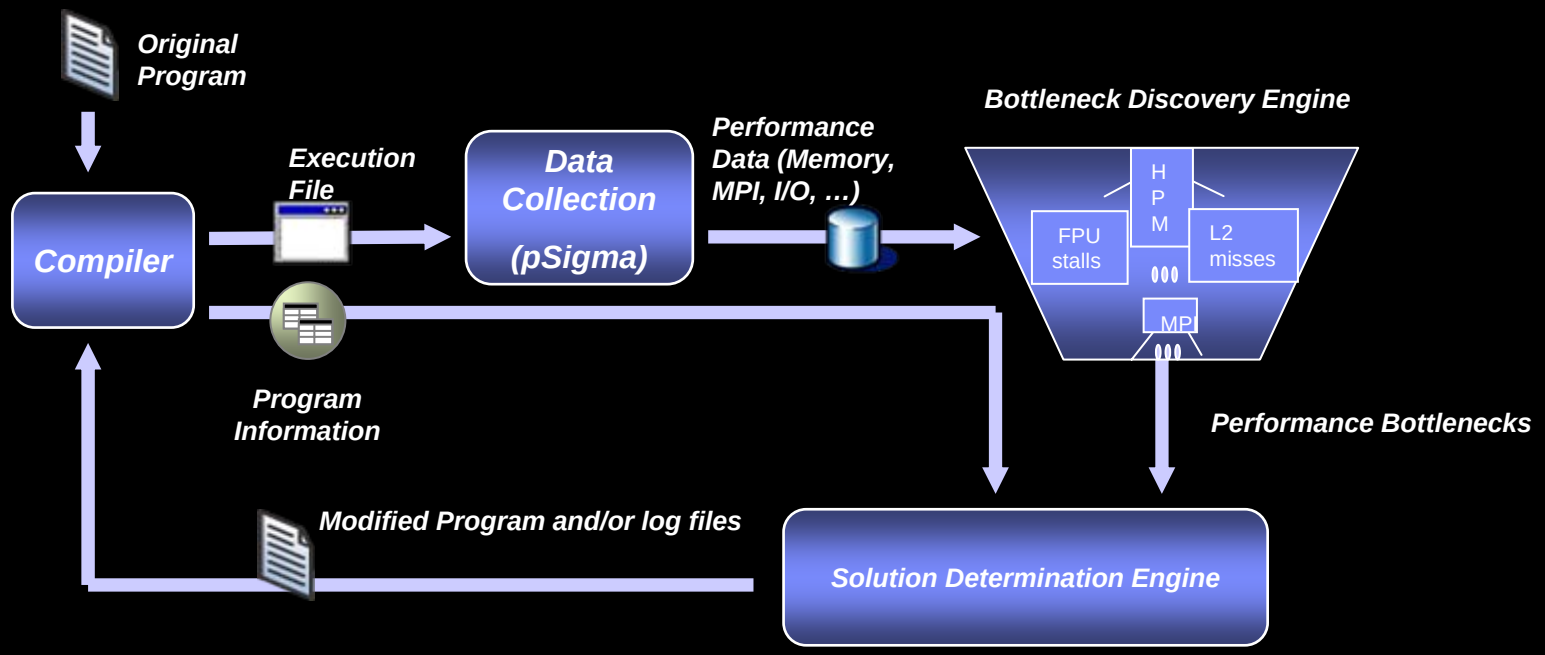
Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

PERCS Impact on Productivity Gap

- **State-of-Art Application Enablement circa 2002+**
  - Source code modification (e.g., timing routines)
  - Non-selective, non-source code correlated tools (e.g., PAPI)
  - Dynamic instrumentation via external agents (e.g., DynInst)
  - GUI frameworks to look at data (e.g., Vampir, Vtune, Tau)
  - No unified analysis framework (CPU, MPI, OpenMP, and I/O)
  - No management of large scale performance data
- **IBM DARPA HPCS Toolkit**
  - Next generation unified framework for automated (not automatic) intelligent-assist of application performance tuning including…
    - No source code modifications…but with source code correlation of the data
    - Selective and dynamic instrumentation without external agents
    - Large scale data management
- **In a Nutshell:**
  - Previous tools only show you the data…does not resolve the Productivity Gap.
  - The HPCS Toolkit makes sense of the data…closes the Productivity Gap.

Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

High Level Design Flow for HPCS Toolkit



- **HPCS Toolkit provides Automated Framework for Performance Analysis.**
  - Intelligent automation of performance evaluation and decision system.
  - Interactive capability with graphical/visual interface always available.

Bottleneck: elapsed time exceeds threshold for completing work.

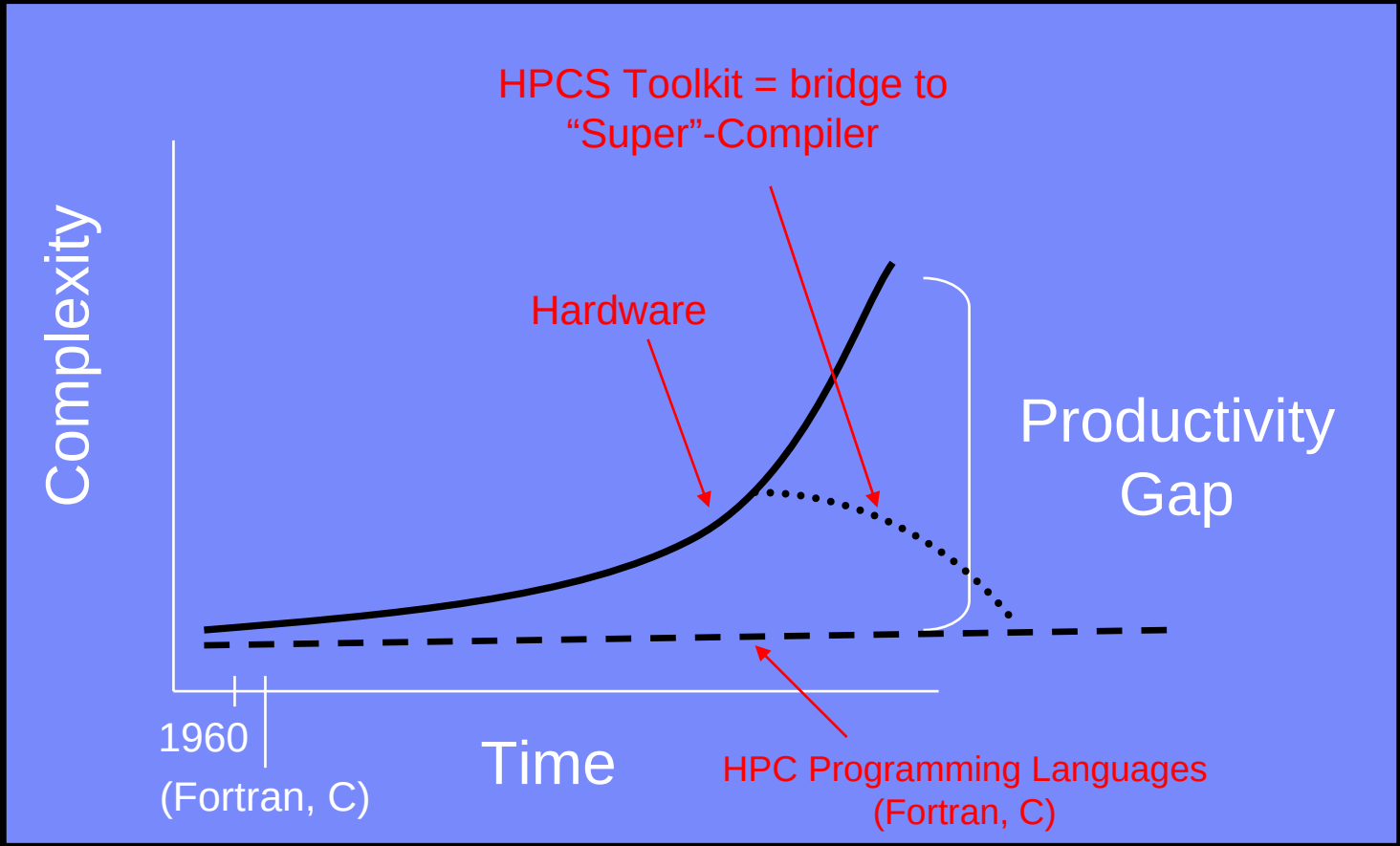
Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

HPCS Toolkit Scalability

- **Self-Contained Performance Data Collection Framework**
  - Part of the instrumented application executable
    - No background processes or external agents
    - Extensible to MRNet (University of Wisconsin)
- **Use of Parallel File System (GPFS)**
  - Data managed in parallel via distributed files
    - Up to five files per process (e.g., for each MPI task):
      1. HPM data
      2. MPI data
      3. OpenMP data
      4. Memory reference data
      5. I/O data
- **Pre-runtime and Post-runtime Filtering Capability**
  - User-defined logic to reduce data to be captured and/or analyzed
- **IBM Research Blue Gene test-bed**
  - Up to 0.5 million processor systems

Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

Closing the Enablement Productivity Gap



Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

Automated Performance Tuning – Timetable

2007 Deliverables:

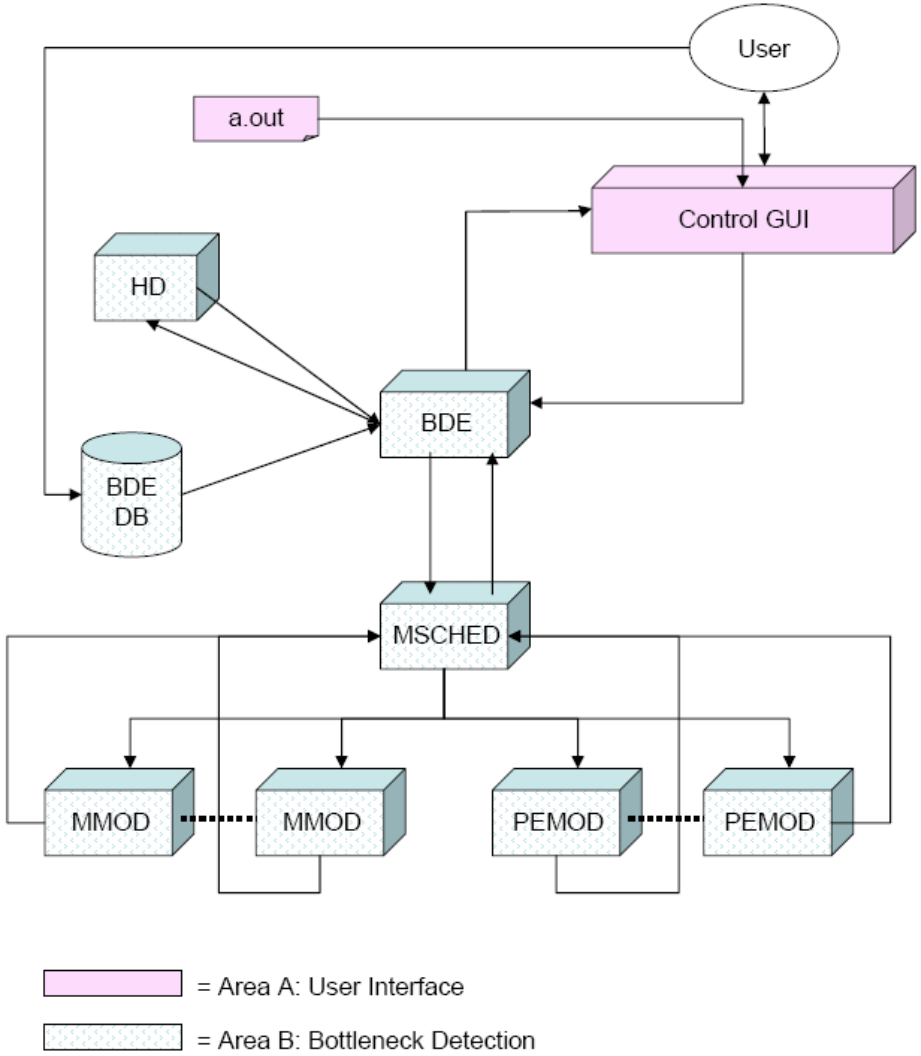
- **Performance Data Collection**
  - Scalable, dynamic, programmable
  - Completely binary: no source code modification to instrument application…
  - But retains ability to correlate all performance data with source code
- **Bottleneck Discovery**
  - Make sense of the performance data
  - Mines the performance data to extract bottlenecks

FUTURE MILESTONE DELIVERABLES:

- **Solution Determination - 2008 - 2009**
  - Make sense of the bottlenecks
  - Mines bottlenecks and suggests system solutions (hardware and/or software)
  - Assist compiler optimization (including custom code transformations)
- **Performance "Visualization" - 2008 - 2010**
  - Performance Data / Bottleneck / Solution Information feedback to User
    - Logging (textual information)
    - Compiler feedback
  - Output to other tools (e.g., Kojak analysis, Paraver visualization, Tau, etc.)

Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

BDE Architecture



Abbreviation	Meaning
BDE	Bottleneck Detection Engine
BDE DB	Bottleneck Detection Engine Database
MMOD	Metric Module
PEMOD	Performance Estimation Module
MSCHED	Performance Scheduler
HD	Hotspot Detector

Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

Example Codes and Applications

- **Standard benchmark codes/kernels**
  - SPEC SWIM (Shallow Water Model)
  - NAS Parallel Bechmarks
- **DARPA Applications**
  - MHD (Magneto Hydrodynamics)
    - Model turbulence and collision in fluids
    - Integration, collision, and stream are time consuming steps
  - HYCOM (Hybrid Ocean Model)
    - 2D ocean circulation model with hybrid coordinate scheme to improve modeling accuracy as the ocean depth varies from deep stratified water to shallow coastal regions

Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

HD Results (Loop Level)

hpcs_gui

File Action Database Window

BDE | BDE and SDE

DATA WINDOW

HD | BDE

	SELECTION	FILE	FUNCTION	START	END L	self secs (P0)
1	☑ Loop	mhd.F	collision	512	592	5.03
2	☐ Function	mhd.F	collision	513	596	5.03
3	☐ Function	mhd.F	stream	338	494	3.92
4	☑ Loop	mhd.F	stream	416	493	2.67
5	☐ Loop	mhd.F	stream	339	365	0.64
6	☐ Loop	mhd.F	stream	367	414	0.61
7	☑ Function	mhd.F	neighbours	619	807	0.22
8	☑ Loop	mhd.F	neighbours	750	804	0.14
9	☑ Loop	mhd.F	equil	291	314	0.13
10	☐ Function	mhd.F	equil	292	317	0.13
11	☐ Loop	mhd.F	check	927	933	0.07
12	☐ Function	mhd.F	check	924	934	0.07
13	☐ Loop	mhd.F	neighbours	693	748	0.03
14	☐ Loop	mhd.F	neighbours	618	626	0.03
15	☐ Function	mhd.F	init	237	265	0.02
16	☐ Loop	mhd.F	neighbours	644	651	0.02
17	☐ Loop	mhd.F	init	273	281	0.02
18	☐ Loop	mhd.F	neighbours	681	689	0
19	☐ Function	mhd.F	my_range	224	226	0
20	☐ Loop	mhd.F	init	264	267	0
21	☐ Loop	mhd.F	init	268	271	0
22	☐ Function	mhd.F	mhd	31	145	0
23	☐ Loop	mhd.F	neighbours	656	664	0
24	☐ Loop	mhd.F	mhd	117	129	0
25	☐ Function	mhd.F	decomp	157	214	0
26	☐ Loop	mhd.F	neighbours	899	906	0

SOURCE WINDOW

mhd.F

```
421     !fetch one line strides = 16 elements at a time
422     ptr => geq(inmp-1, j, 1, k)
423     CALL Mambo_Prefetch(ptr,   128, 1, NMP_PREF_LINES)
424     ptr => geq(inmp, j-1, 3, k)
425     CALL Mambo_Prefetch(ptr,   128, 1, NMP_PREF_LINES)
426     ptr => geq(inmp+1, j, 5, k)
427     CALL Mambo_Prefetch(ptr,   128, 1, NMP_PREF_LINES)
428     ptr => geq(inmp, j+1, 7, k)
429     CALL Mambo_Prefetch(ptr,   128, 1, NMP_PREF_LINES)
430     do i = inmp, MIN(inmp+16*NMP_PREF_LINES-1, iend)
431     #else
432     do i = ista, iend
433     #endif
434       g(i,j,1,k)= geq(i-1,j,1,k)
435       g(i,j,3,k)= geq(i,j-1,3,k)
436       g(i,j,5,k)= geq(i+1,j,5,k)
437       g(i,j,7,k)= geq(i,j+1,7,k)
438     enddo
439     #ifdef NMP
440     enddo
441     #endif
442     enddo
443
444
445     do j = jsta, jend
```


Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

Source Code with Clock Ticks

hpcs_gui

File Action Database Window

BDE | BDE and SDE

mhd.F

```
347          ptr => feq(inmp, j-1, 3)
348          CALL Mambo_Prefetch(ptr,   128, 1, NMP_PREF_LINES)
349          ptr => feq(inmp+1, j, 5)
350          CALL Mambo_Prefetch(ptr,   128, 1, NMP_PREF_LINES)
351          ptr => feq(inmp, j+1, 7)
352          CALL Mambo_Prefetch(ptr,   128, 1, NMP_PREF_LINES)
353          do i = inmp, MIN(inmp+(16*NMP_PREF_LINES)-1, iend)
354          #else
355          do i = ista, iend
356          #endif
357          !  write(*,*) i, '   ', j, '   ', inmp
358     21     f(i,j,1)= feq(i-1,j,1)
359     28     f(i,j,3)= feq(i,j-1,3)
360      9    f(i,j,5)= feq(i+1,j,5)
361      6    f(i,j,7)= feq(i,j+1,7)
362          enddo
363          #ifdef NMP
364          enddo
365          #endif
366          enddo
367
368          do j = jsta, jend
369          #ifdef NMP
370          do inmp = ista, iend-16*NMP_PREF_LINES+1, 16*NMP_PREF_LINES
371          !fetch one line strides = 16 elements at a time
372          ptr => feq(inmp, j, 2)
373          CALL Mambo_Prefetch(ptr,   128, 1, NMP_PREF_LINES)
```


Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

BDE Results

hpcs_gui

File Action Database Window

SDE

HD | BDE

	BOTTLENEC	DIMENS	DESCRIPTION	RULE	FUNCTION	START LI	END LINE
1	DcacheMiss	CPU	cycles wasted due to stalls	(PM_RUN_CYC - (PM_GCT_EMPTY_CYC + PM_GRP_CMPL))	stream	416	493
2	DcacheMiss	CPU	cycles wasted due to stalls	(PM_RUN_CYC - (PM_GCT_EMPTY_CYC + PM_GRP_CMPL))	equil	291	314
3	DcacheMiss	CPU	cycles wasted due to stalls	(PM_RUN_CYC - (PM_GCT_EMPTY_CYC + PM_GRP_CMPL))	neighbours	750	804
4	DcacheMiss	CPU	cycles wasted due to stalls	(PM_RUN_CYC - (PM_GCT_EMPTY_CYC + PM_GRP_CMPL))	collision	512	592
5	DcacheMiss	CPU	cycles wasted due to stalls	(PM_RUN_CYC - (PM_GCT_EMPTY_CYC + PM_GRP_CMPL))	neighbours	619	807
6	MPIBarrier1	COMMU	imbalanced wait time	MAX(mpi_wait_hot_sum_time)-MIN(mpi_wait_hot_sum_time) > 1.	neighbours	619	807
7	Stalls	CPU	cycles spent in stalls	(PM_CMPLU_STALL_LSU + PM_CMPLU_STALL_FXU + PM_CM	neighbours	619	807
8	Stalls	CPU	cycles spent in stalls	(PM_CMPLU_STALL_LSU + PM_CMPLU_STALL_FXU + PM_CM	stream	416	493
9	Stalls	CPU	cycles spent in stalls	(PM_CMPLU_STALL_LSU + PM_CMPLU_STALL_FXU + PM_CM	collision	512	592
10	Stalls	CPU	cycles spent in stalls	(PM_CMPLU_STALL_LSU + PM_CMPLU_STALL_FXU + PM_CM	neighbours	750	804
11	Stalls	CPU	cycles spent in stalls	(PM_CMPLU_STALL_LSU + PM_CMPLU_STALL_FXU + PM_CM	equil	291	314

Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

Query Execution

hpcs_gui <2>

- Fri Apr 11 15:36:43 2008
  - /gsa/yktgsa-h1/01/hfwen/test/DARPA/mhd.AIX/mhdNopg
    - mpi
      - mhd.F:collision:512:592
      - mhd.F:equil:291:314
      - mhd.F:neighbours:619:807

Label	metric	Value for P0	MIN	MAX	AVG
	mpi_all_hot_avg_time	0.000772	0.000772	0.000954	0.000439
	mpi_all_hot_max_time	0.079773	0.079773	0.078625	0.042089
	mpi_all_hot_med_time	0.000005	0.000005	0.000005	0.000007
	mpi_all_hot_min_time	0.000001	0.000001	0.000001	0.000001
	mpi_all_hot_sum_time	3.089104	3.089104	3.816296	1.757381
	mpi_all_hot_var_time	0.000055	0.000055	0.000068	0.000031
	mpi_all_prog_avg_time	0.000955	0.000955	0.000946	0.000483
	mpi_all_prog_max_time	0.079773	0.079773	0.078625	0.042089
	mpi_all_prog_med_time	0.000005	0.000005	0.000005	0.000007
	mpi_all_prog_min_time	0.000001	0.000001	0.000001	0.000001
	mpi_all_prog_sum_time	3.854623	3.854623	3.817104	1.949907
	mpi_all_prog_var_time	0.000069	0.000069	0.000067	0.000034
	mpi_wait_hot_avg_time	0.001535	0.001535	0.001900	0.000864
	mpi_wait_hot_max_time	0.079773	0.079773	0.078625	0.042089
	mpi_wait_hot_med_time	0.000002	0.000002	0.000003	0.000003
	mpi_wait_hot_min_time	0.000001	0.000001	0.000001	0.000001
	mpi_wait_hot_sum_time	3.069639	3.069639	3.800613	1.727596
	mpi_wait_hot_var_time	0.000110	0.000110	0.000134	0.000061
	mpi_wait_prog_avg_time	0.001535	0.001535	0.001900	0.000864
	mpi_wait_prog_max_time	0.079773	0.079773	0.078625	0.042089
	mpi_wait_prog_med_time	0.000002	0.000002	0.000003	0.000003
	mpi_wait_prog_min_time	0.000001	0.000001	0.000001	0.000001
	mpi_wait_prog_sum_time	3.069639	3.069639	3.800613	1.727596
	mpi_wait_prog_var_time	0.000110	0.000110	0.000134	0.000061

      - mhd.F:neighbours:750:804
      - mhd.F:stream:416:493
    - pomprof
- Mon Apr 7 11:27:04 2008

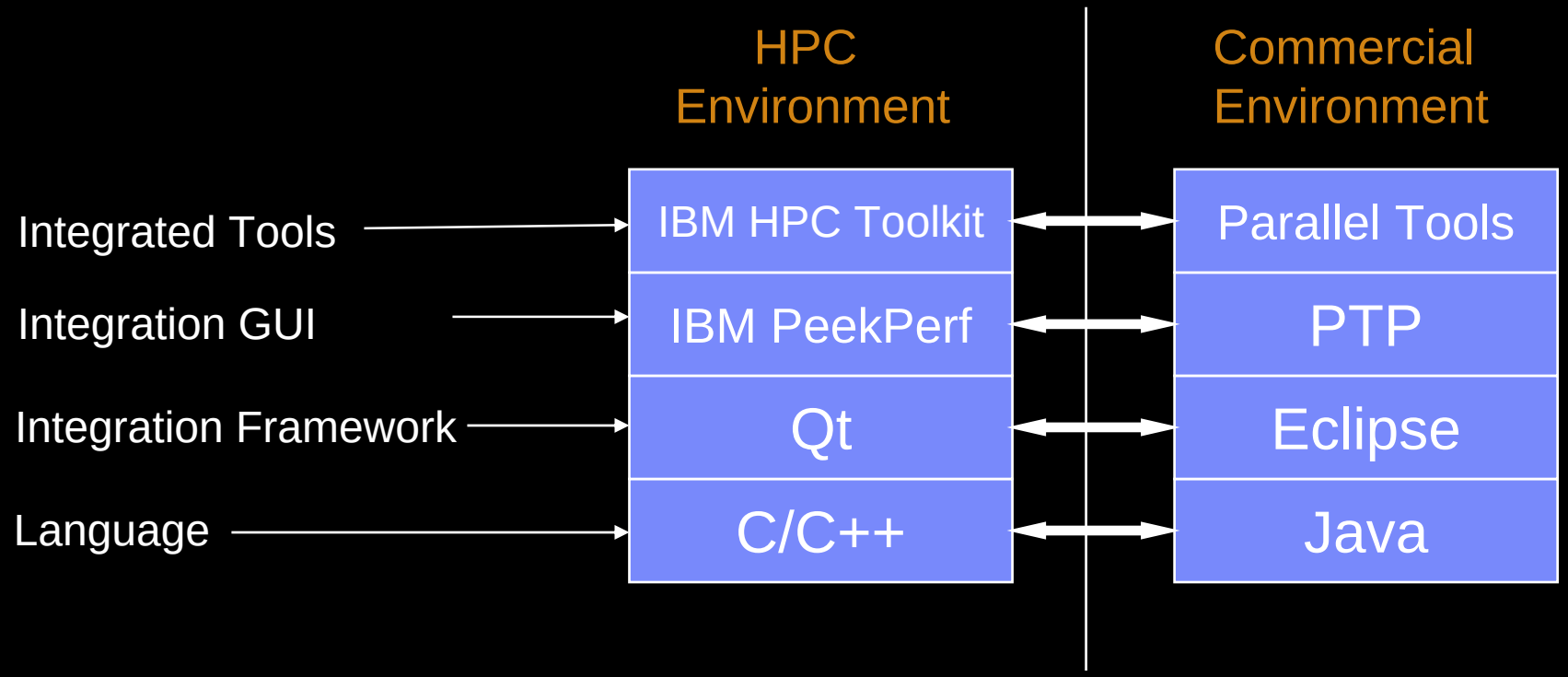
Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

HPCS Toolkit Platform Availability

- **Application programming languages supported**
  - Fortran, C/C++
  - On Horizon: UPC, CoA Fortran, X10 ?
- **GUI environments supported**
  - Qt (Nokia/Trolltech) - C/C++ based, open source
  - Eclipse (IBM) - Java based, open source
- **Hardware platforms supported**
  - PERCS and IBM POWER servers
  - IBM Blue Gene systems
  - On Horizon: AMD and Intel-based machines, Cell, ?

Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

Commercialization: Migration to Eclipse Platform



- Both Qt and Eclipse are Open Source and have similar functionalities.
- Eclipse is funded by the DARPA HPCS Project and is popular in Commercial environment (Rational Development tools)
- Qt is much more popular in HPC environments (e.g., Linux standard desktop GUI)

Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

HPCST is Ongoing Research

- **We welcome collaborators and early adopters to help grow BDE DB and provide more feedback on real applications.**
- **We also have post-doc, co-op and student internships available at the IBM Watson lab!**

Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.
Data Marking: OTHER DATA

Thank you!

Use, reproduction, or disclosure is subject to the restrictions as stated in Agreement #HR0011-07-9-0002 between the Government and IBM.
May not be disclosed except in accordance with 5 CFR 552, 18 USC 1905, 18 USC 1831-2, 18 USC 1838, 41 USC 423, FAR Part 9.505-4 and other applicable federal law and regulation.

Data Marking: OTHER DATA