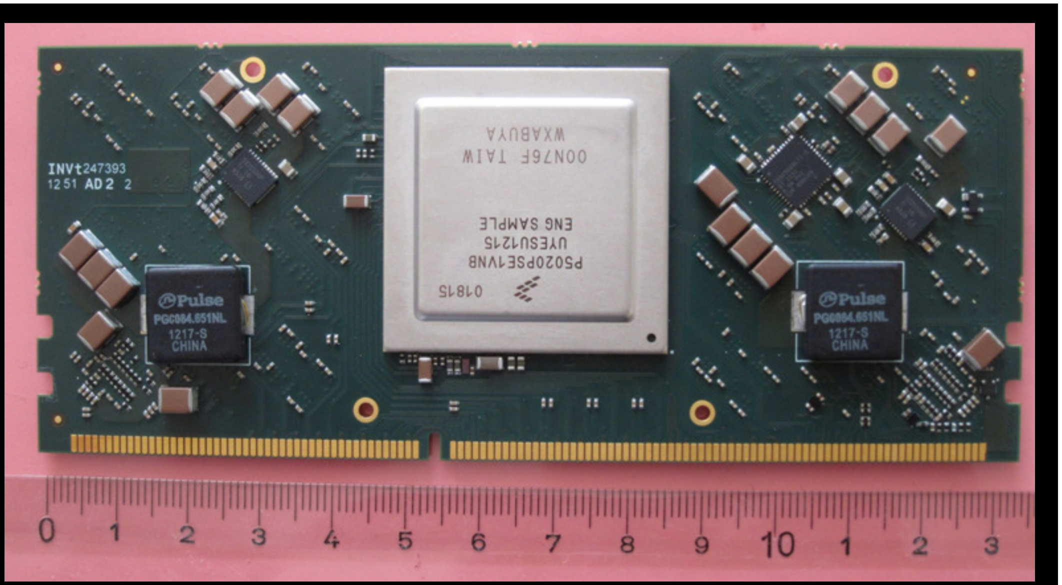
http://www.swissdutch.ch:6999/



The IBM-DOME microserver demonstrator

Ronald P. Luijten – Data Motion architect

1 May 2013

DOME – Research Phase for SKA (SKA = Square Kilometer Array)

The SKA will be the largest and most sensitive radio telescope ever built.

A single instrument with >10’000’s of antennas will become operational in 2024 with frequency ranges 70MHz to 10GHz. This will generate huge amounts of data, which need to be *transported, analyzed, stored and retrieved* – at *very low* power and *very low* cost.

A true Exascale Analytics Challenge!

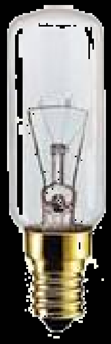
DOME is a research phase project before start of SKA deployment in 2017

- 5 year collaboration between ASTRON (NL) and IBM, started Feb 2012
- Co-funded by Dutch government and IBM
- Multi project program including high scale-out and scale-in micro server project

IBM DOME µServer Motivation & Objectives

- **Create *the* worlds highest density 64 bit µ-server drawer**
  - Useful for both SKA radioastronomy and IBM future capability
  - Very high energy efficiency

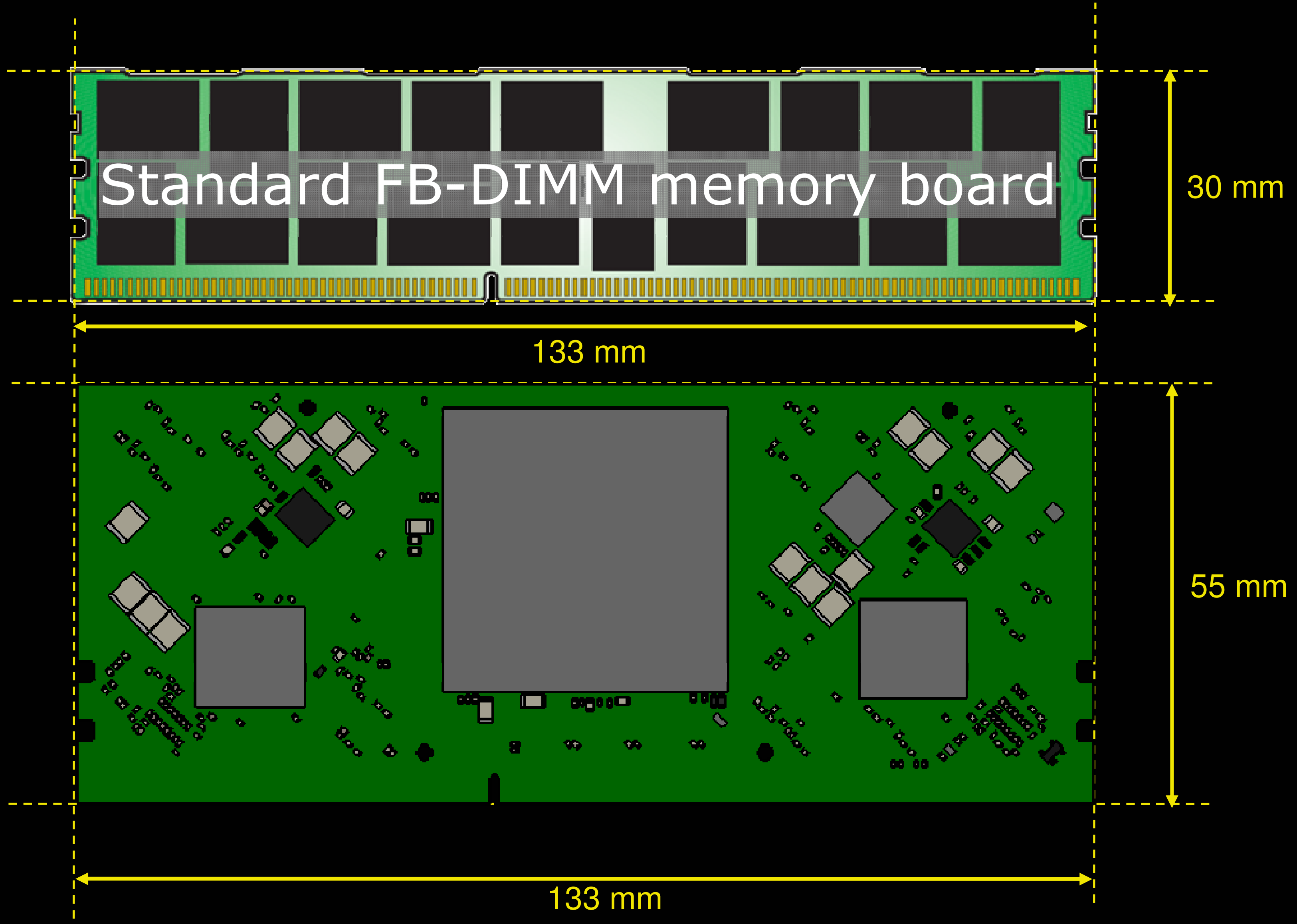
- **Most efficient cooling using IBM technology (ref: SuperMUC TOP500 pos #4)**
- Platform for Business Analytics appliance pre-product research
- "Datacenter in-a-box"



- Must be true 64 bit to enable business applications
  - Currently precludes ARM (currently no 64-bit Silicon available)
  - PPC64 is most compelling based on ecosystem compatibility
- Must run server class OS (SLES11 or RHEL6, or equivalent)
- Must use commodity components only, HW standards, standard SW based
- Must be a true microserver (IBM ZRL definition):
  - integrates the entire compute node motherboard, except DRAM and NOR-boot flash
  - Must integrate Ethernet on 'microserver' SOC.

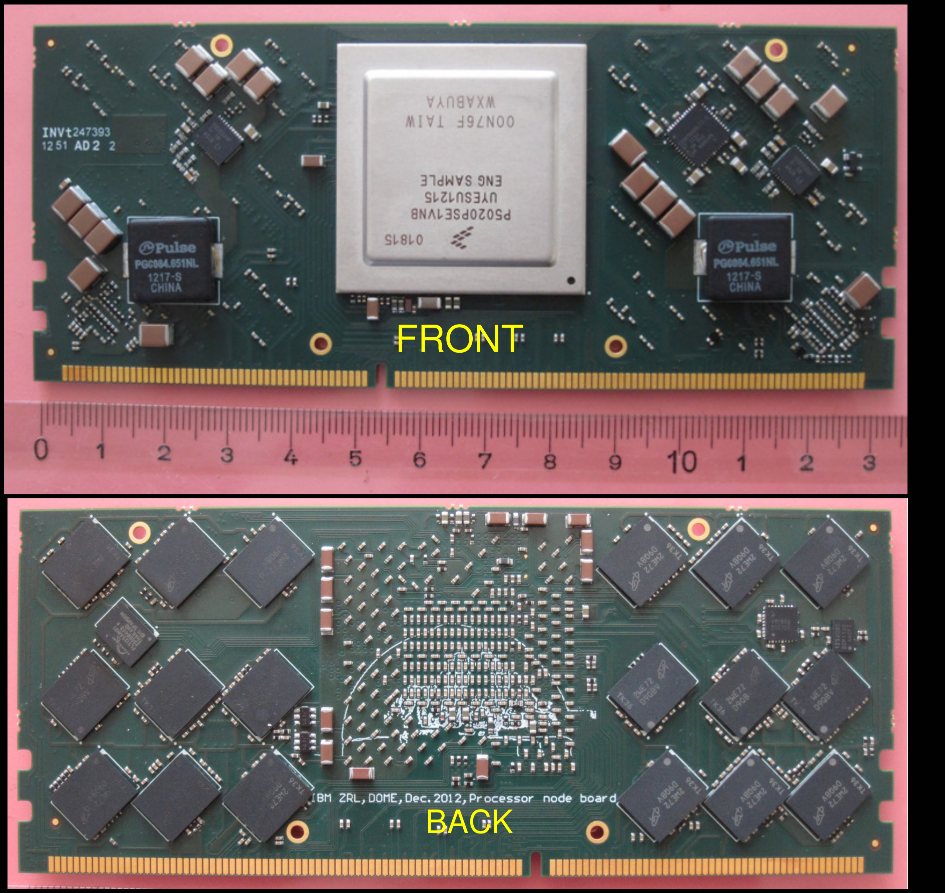
- **This is a research project – capability demonstrator only**

DOME Demonstrator compute node board

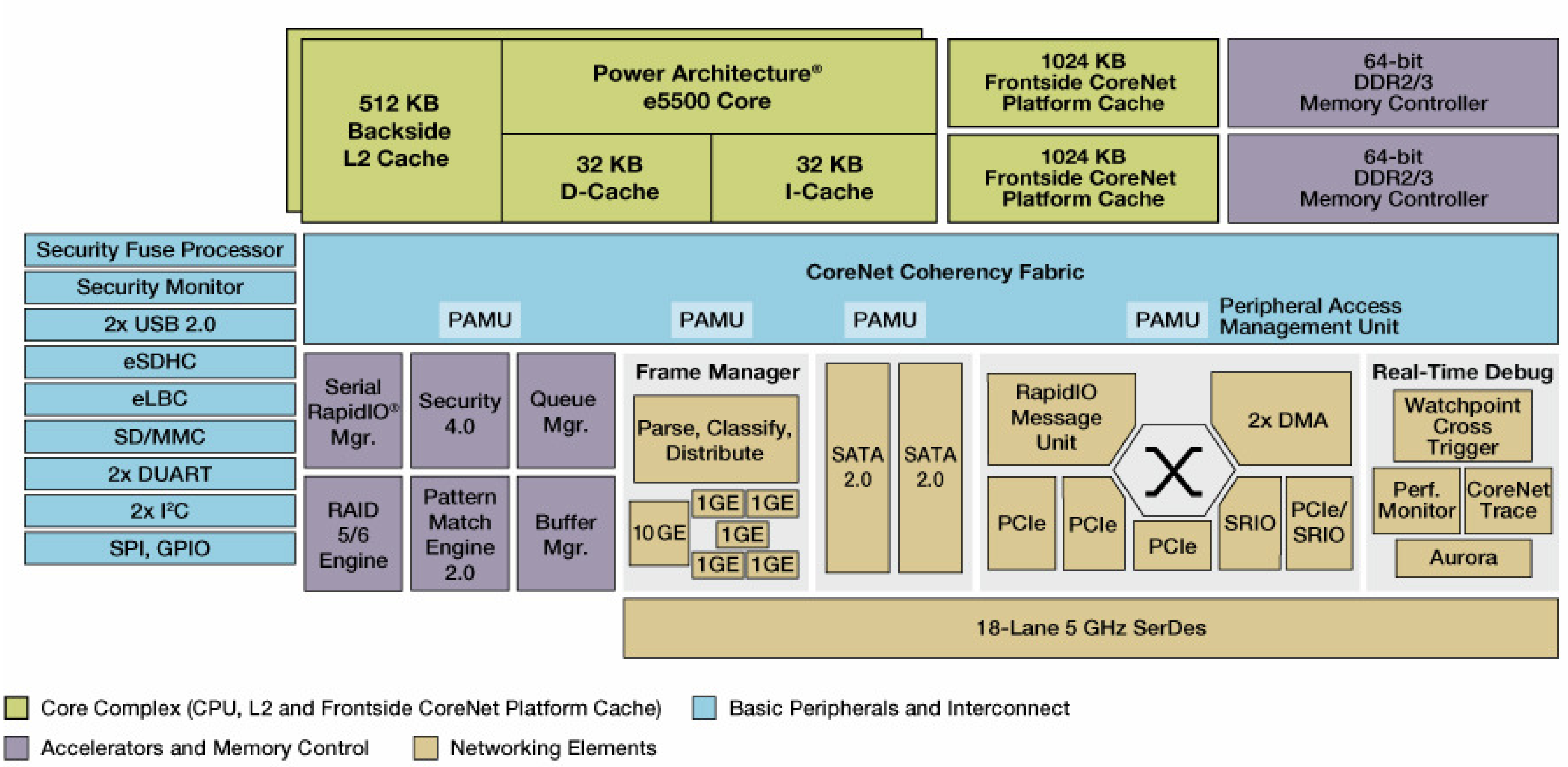


Compute node interfaces across DIMM connector

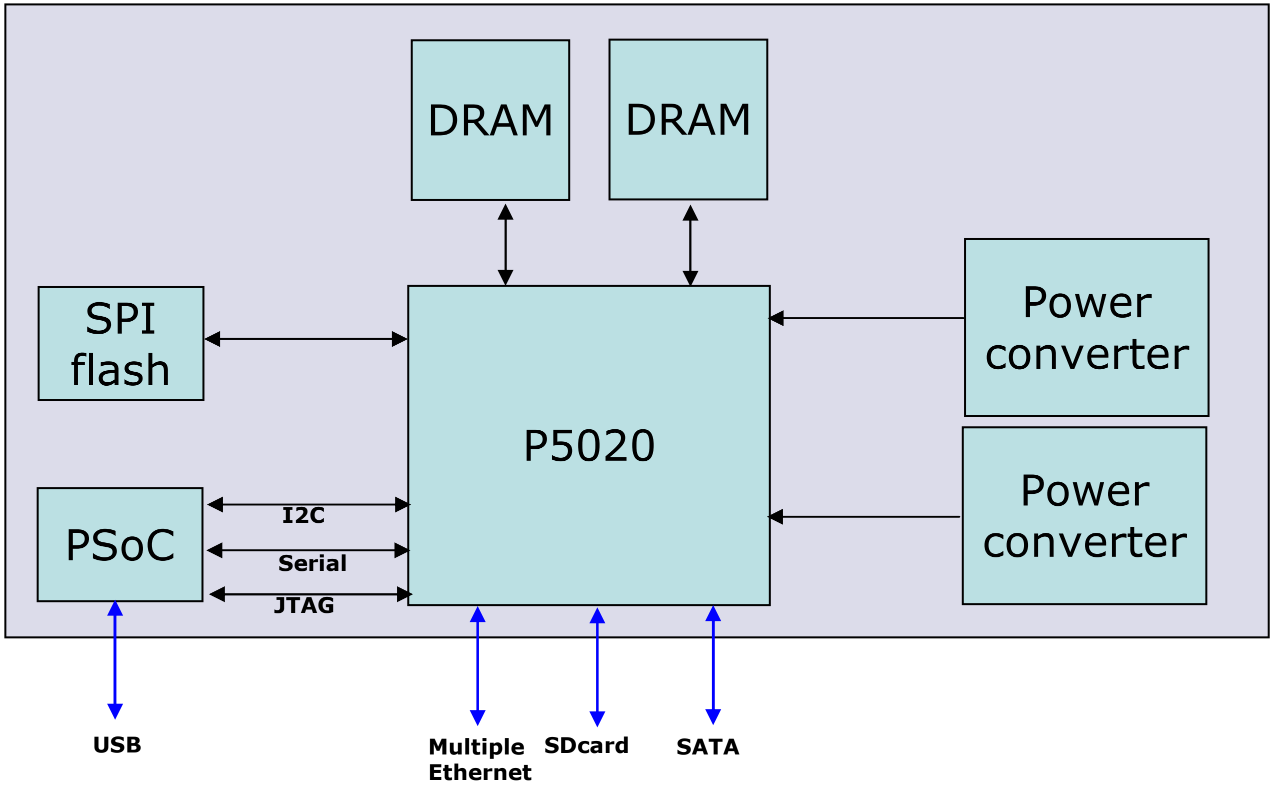
- 1 interface SATA
- 5 interfaces Gigabit ethernet
- 2 interfaces 10 Gigabit ethernet
- SD card interface
- USB interface
- Various power supply levels



FSL P5020 SoC block diagram



IBM / ASTRON compute node board diagram



And now the Software story…

And now the Software story…

NOTE

The Freescale 64 bit PowerPC parts are using the latest PPC64 architecture

However, a key difference with IBMs PPC (eg. P7) is that FSL builds

BOOK-E

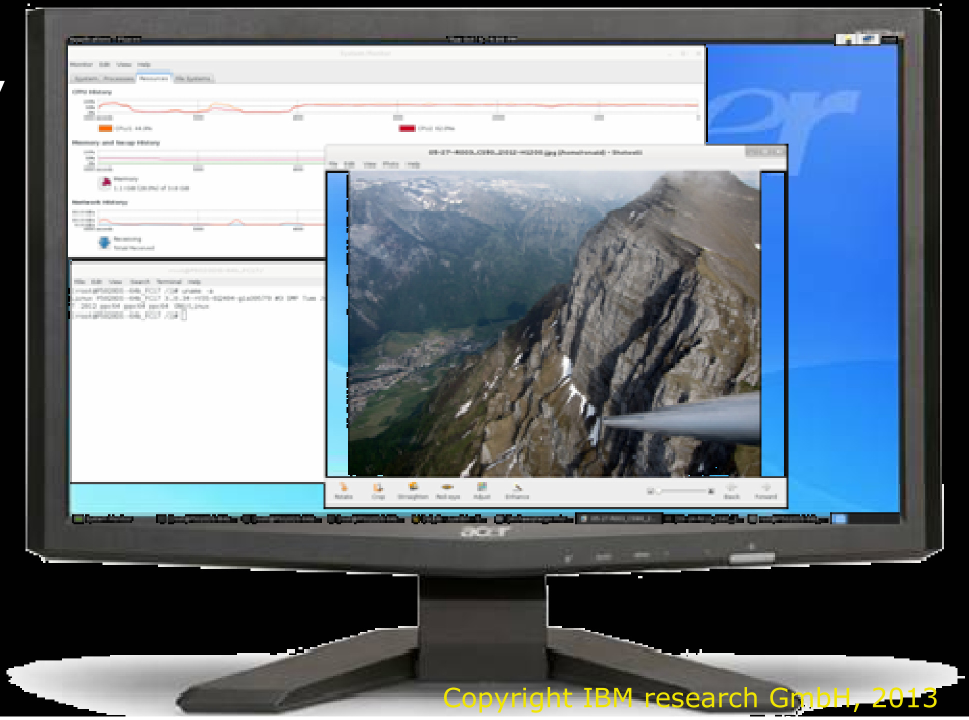
Which causes SW challenges. The next few foils show how we overcome this

64 bit Fedora 17 on P5020DS

- Freescale took kernel version 3.0.34 from kernel.org
- Configured and compiled it for P5020
- Took Fedora user space root FS (thru another PPC platform)
- Runs 100% OK - YUM, Gnome desktop, networking, apache, etc…
  - System up and running > 40 days
  - Java, Python, …

- This effort took approximately ONE day

✓

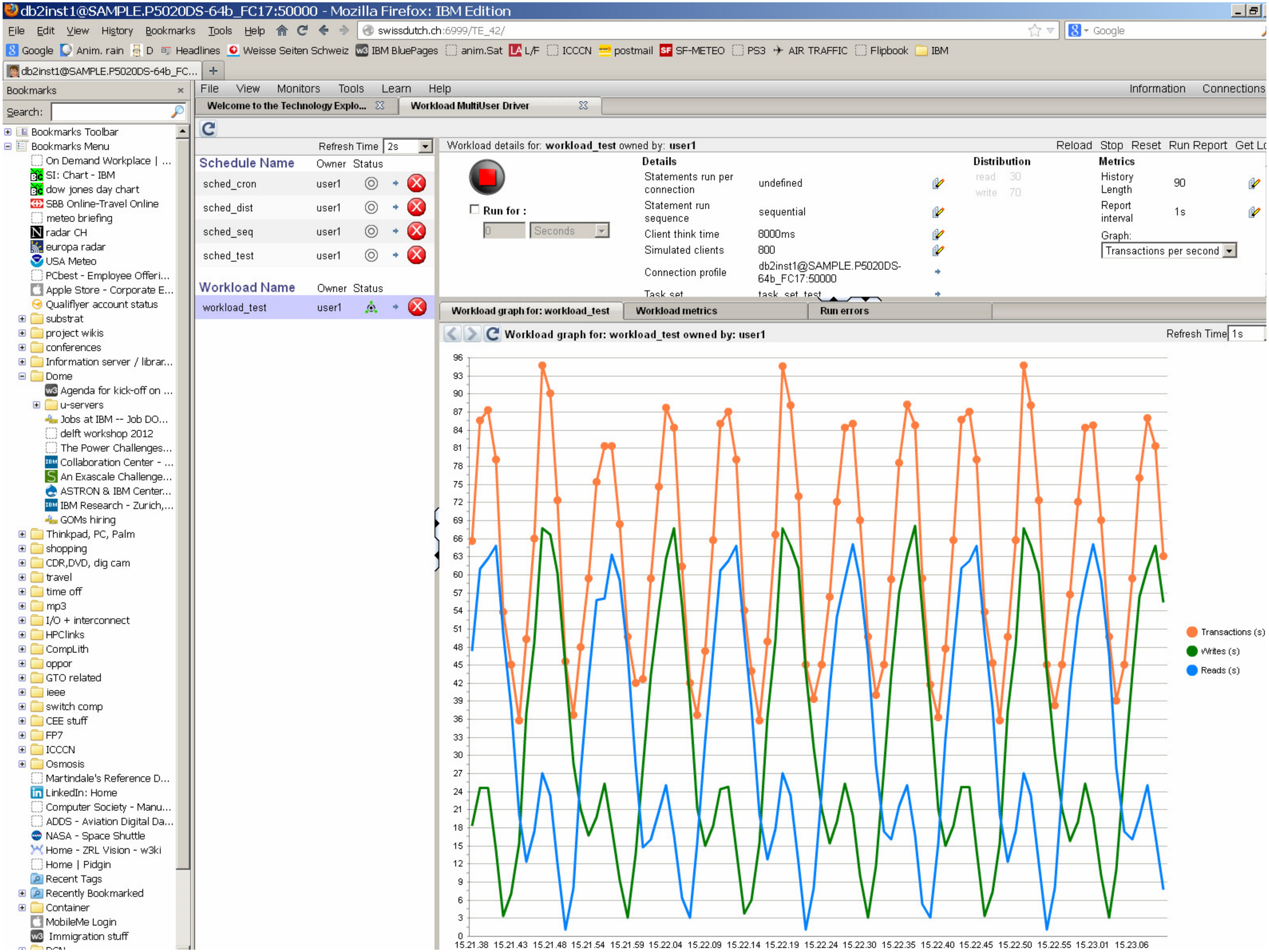


IBM DB2 installation on P5020

- Simple install of IBM XL C/C++ runtime (XLC compiler runtime)
- Install libaio
- Simple install of IBM DB2 (express-C, v10.1)
- Some minor configuration adjustments required
- Entire process only took a few hours -- *no compilation was needed*
- Demo available
  - Technology explorer (runs php in browser)
  - WMD Workload Multi-User Driver (Java based)
  - DB2 data base engine

- Runs stable – able to exercise without any issues

✓



Schedule Name	Owner	Status
sched_cron	user1	
sched_dist	user1	
sched_seq	user1	
sched_test	user1	

Workload Name	Owner	Status
workload_test	user1	

Workload details for: **workload_test** owned by: **user1**

Details	
Statements run per connection	undefined
Statement run sequence	sequential
Client think time	8000ms
Simulated clients	800
Connection profile	db2inst1@SAMPLE.P5020DS-64b_FC17:50000
Task set	task_set_test

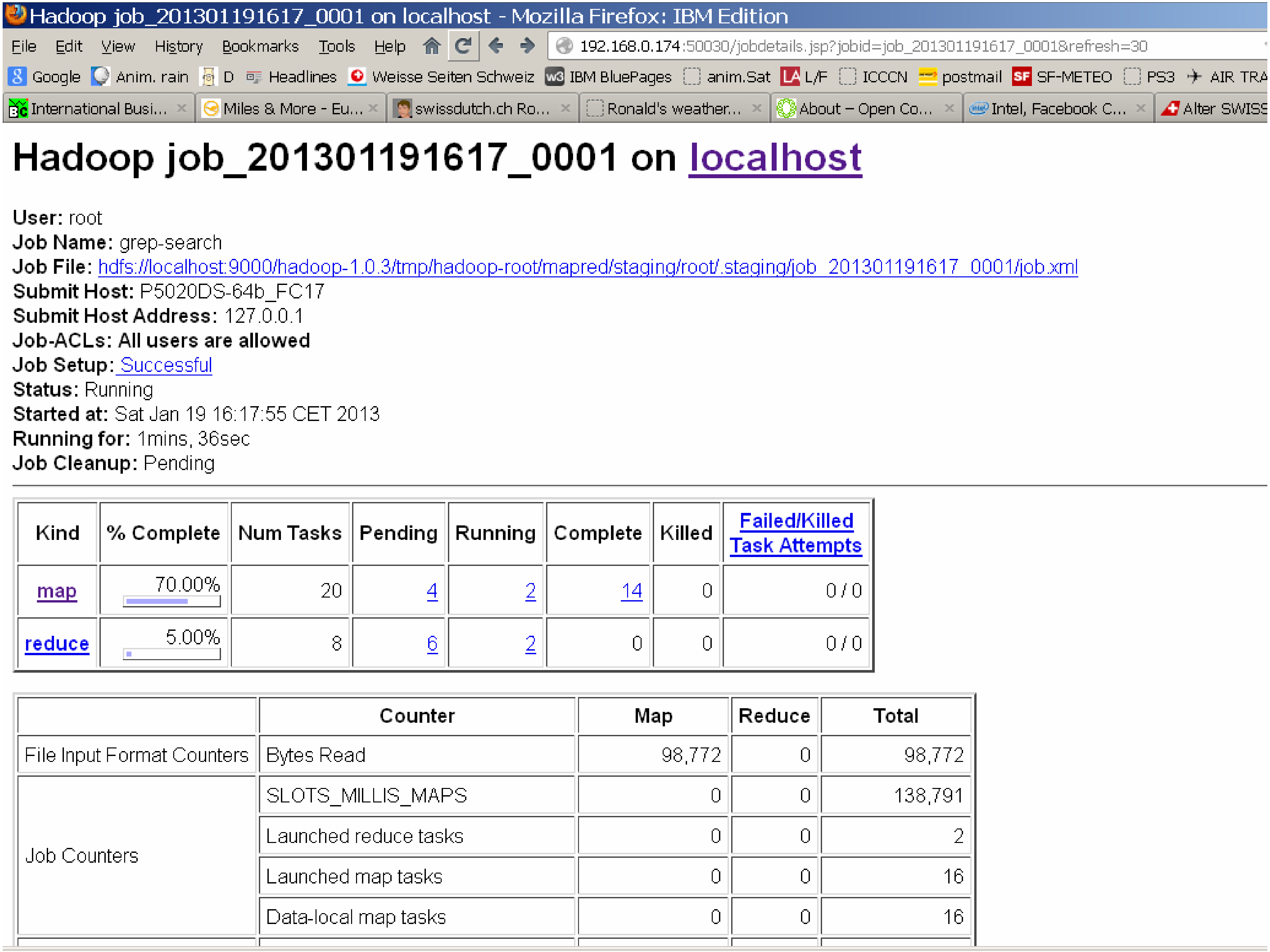
Metrics	
History Length	90
Report interval	1s
Graph:	Transactions per second

Workload graph for: workload_test owned by: user1

Hadoop install on P5020

- Simple install (version 1.0.3 for ppc64)
- Minor configuration effort required
- Works for single node and pseudo-distributed mode
- No compilation necessary
- Demo available

✓ hadoop


Hadoop job_201301191617_0001 on localhost - Mozilla Firefox: IBM Edition
File Edit View History Bookmarks Tools Help
192.168.0.174:50030/jobdetails.jsp?jobid=job_201301191617_0001&refresh=30
Google Anim. rain D Headlines Weisse Seiten Schweiz IBM BluePages anim.Sat L/F ICCCN postmail SF-METEO PS3 AIR TRA
International Busi... Miles & More - Eu... swissdutch.ch Ro... Ronald's weather... About – Open Co... Intel, Facebook C... Alter SWISS

Hadoop job_201301191617_0001 on localhost

User: root
Job Name: grep-search
Job File: hdfs://localhost:9000/hadoop-1.0.3/tmp/hadoop-root/mapred/staging/root/.staging/job_201301191617_0001/job.xml
Submit Host: P5020DS-64b_FC17
Submit Host Address: 127.0.0.1
Job-ACLs: All users are allowed
Job Setup: Successful
Status: Running
Started at: Sat Jan 19 16:17:55 CET 2013
Running for: 1mins, 36sec
Job Cleanup: Pending

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	70.00%	20	4	2	14	0	0 / 0
reduce	5.00%	8	6	2	0	0	0 / 0

	Counter	Map	Reduce	Total
File Input Format Counters	Bytes Read	98,772	0	98,772
Job Counters	SLOTS_MILLIS_MAPS	0	0	138,791
	Launched reduce tasks	0	0	2
	Launched map tasks	0	0	16
	Data-local map tasks	0	0	16

HPC CPMD application port

HPC Carr-Parinello Molecular Dynamics package
For Ab Initio simulations - a key HPC application

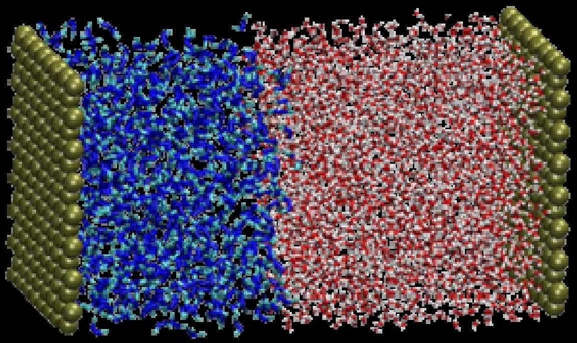
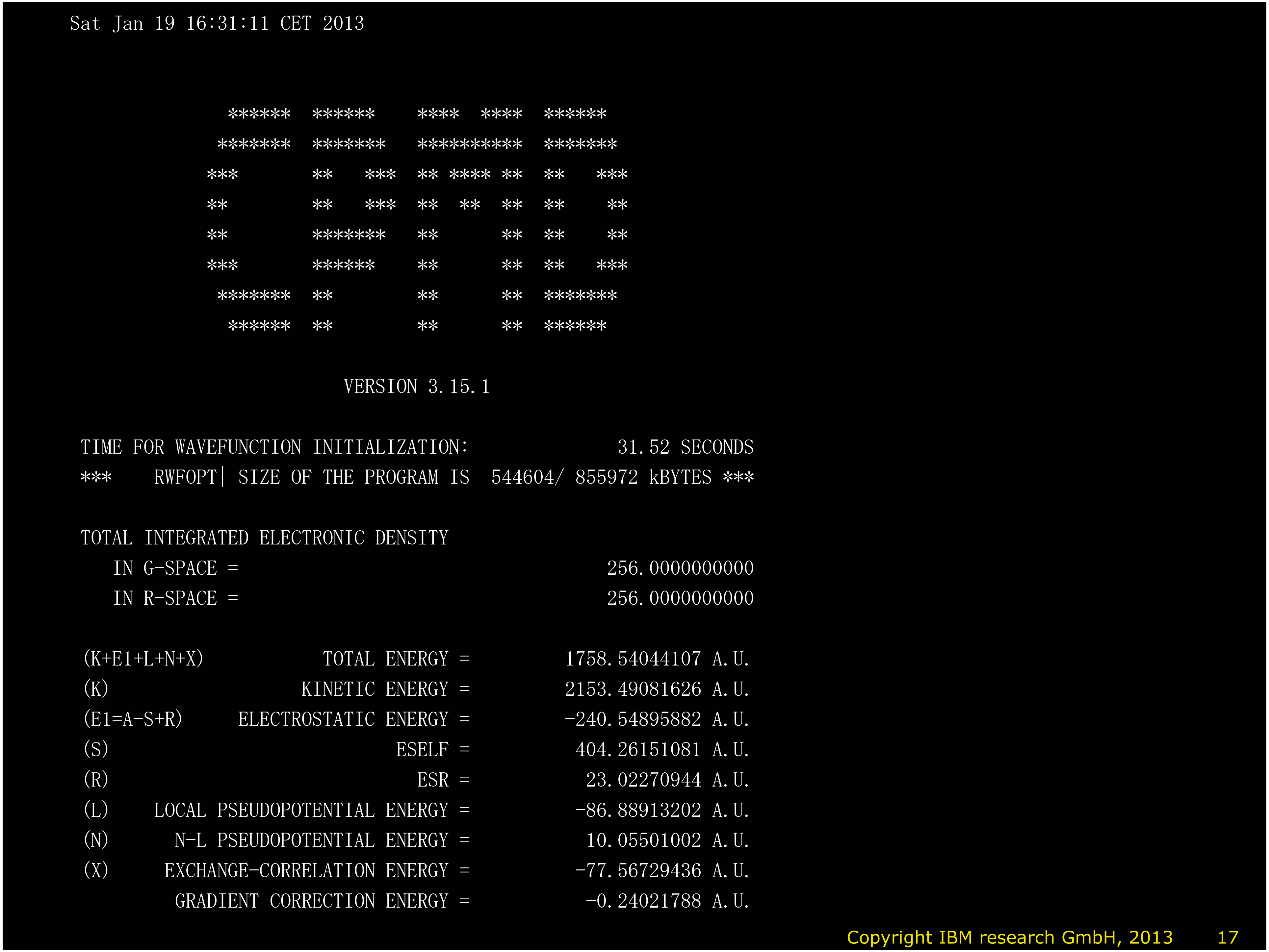


Image Courtesy Jülich Forschungszentrum

- LAPACK install: compile required - 10 min job
  - Using Gfortran and GCC – no errors
- CPMD code base configured for PPC64, 2 cores
  - Natively compiled in 15 mins
  - ~100k lines of Fortran

- Demo available

✓ μCPMD



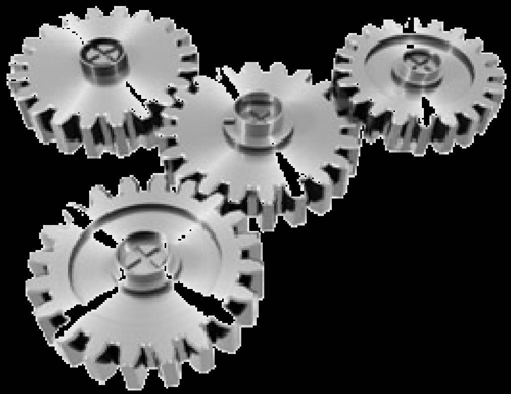
```
Sat Jan 19 16:31:11 CET 2013



               ******  ******    ****  ****  ******
              *******  *******   **********  *******
             ***       **   ***  ** **** **  **   ***
             **        **   ***  **  **  **  **    **
             **        *******   **      **  **    **
             ***       ******    **      **  **   ***
              *******  **        **      **  *******
               ******  **        **      **  ******

                          VERSION 3.15.1

 TIME FOR WAVEFUNCTION INITIALIZATION:              31.52 SECONDS
 ***    RWFOPT| SIZE OF THE PROGRAM IS  544604/ 855972 kBYTES ***

 TOTAL INTEGRATED ELECTRONIC DENSITY
    IN G-SPACE =                                   256.0000000000
    IN R-SPACE =                                   256.0000000000

 (K+E1+L+N+X)           TOTAL ENERGY =         1758.54044107 A.U.
 (K)                  KINETIC ENERGY =         2153.49081626 A.U.
 (E1=A-S+R)     ELECTROSTATIC ENERGY =         -240.54895882 A.U.
 (S)                           ESELF =          404.26151081 A.U.
 (R)                             ESR =           23.02270944 A.U.
 (L)    LOCAL PSEUDOPOTENTIAL ENERGY =          -86.88913202 A.U.
 (N)      N-L PSEUDOPOTENTIAL ENERGY =           10.05501002 A.U.
 (X)     EXCHANGE-CORRELATION ENERGY =          -77.56729436 A.U.
          GRADIENT CORRECTION ENERGY =           -0.24021788 A.U.
```

Conclusion

- Server Class 64 bit OS on PowerPC commodity SOC has arrived

- IBM and Freescale demonstrated on PPC64 Book-E:
  - 64 bit Fedora 17
  - IBM DB2 – no compilation necessary to run
  - Hadoop – no compilation necessary to run
  - HPC CPMD application – straightforward port in a few hours

✓



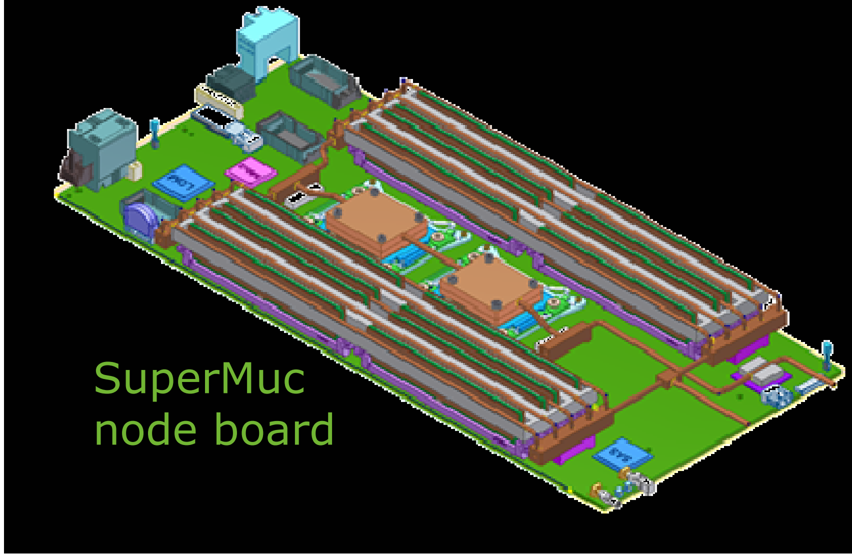
Hot Water Cooling

Most Energy Efficient solution:

- Low PUE possible (<=1.1) – Green IT
- 40% less energy consumption compared to air-cooled systems
- 90% of waste heat can be reused (CO_2 neutral according Kyoto protocol)
- Allows very high density
- Less thermal cycling - improved reliability
- Lower T_j reduces leakage current – further saving energy

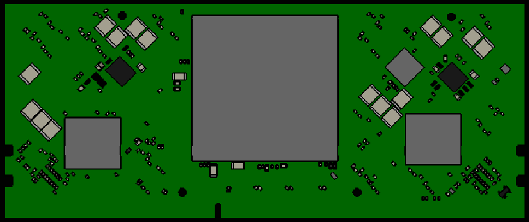
SuperMUC HPC machine at LRZ in Germany demonstrates ZRL hot water cooling

- No 4 on June 2012 TOP500 HPC list

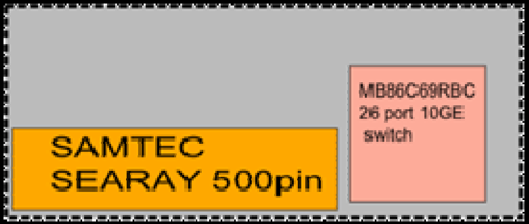


SuperMuc node board

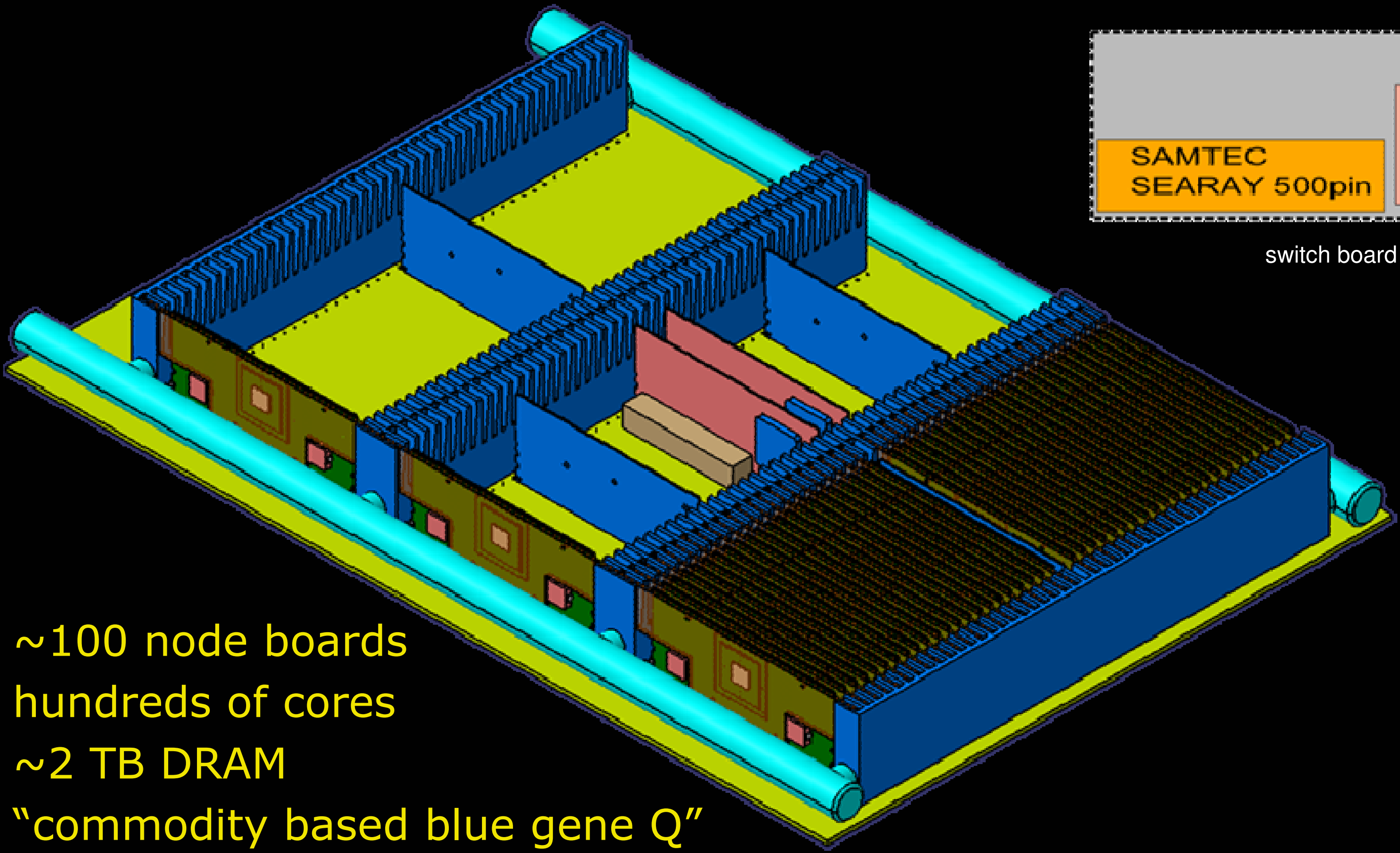
19” 2U Chassis with Combined Cooling and Power



Node board



switch board



~100 node boards

hundreds of cores

~2 TB DRAM

“commodity based blue gene Q”

Status of 1 May 2013

Project start: Feb 2012 (DOME contract signed w/ Dutch government)

Freescale P5040 SoC selected

Freescale relationship established

64 bit, Fedora 17 based Stack running on FSL, Embedded PPC64, BookE – P5020DS

IBM DB2, Hadoop, CPMD

Same SW stack demonstrated at Austin FSL lab on T4240 SoC

First P5020 DOME node board received Feb 2013 – currently in bringup

First 8 way cluster, validating cooling concept, planned 2Q 2013

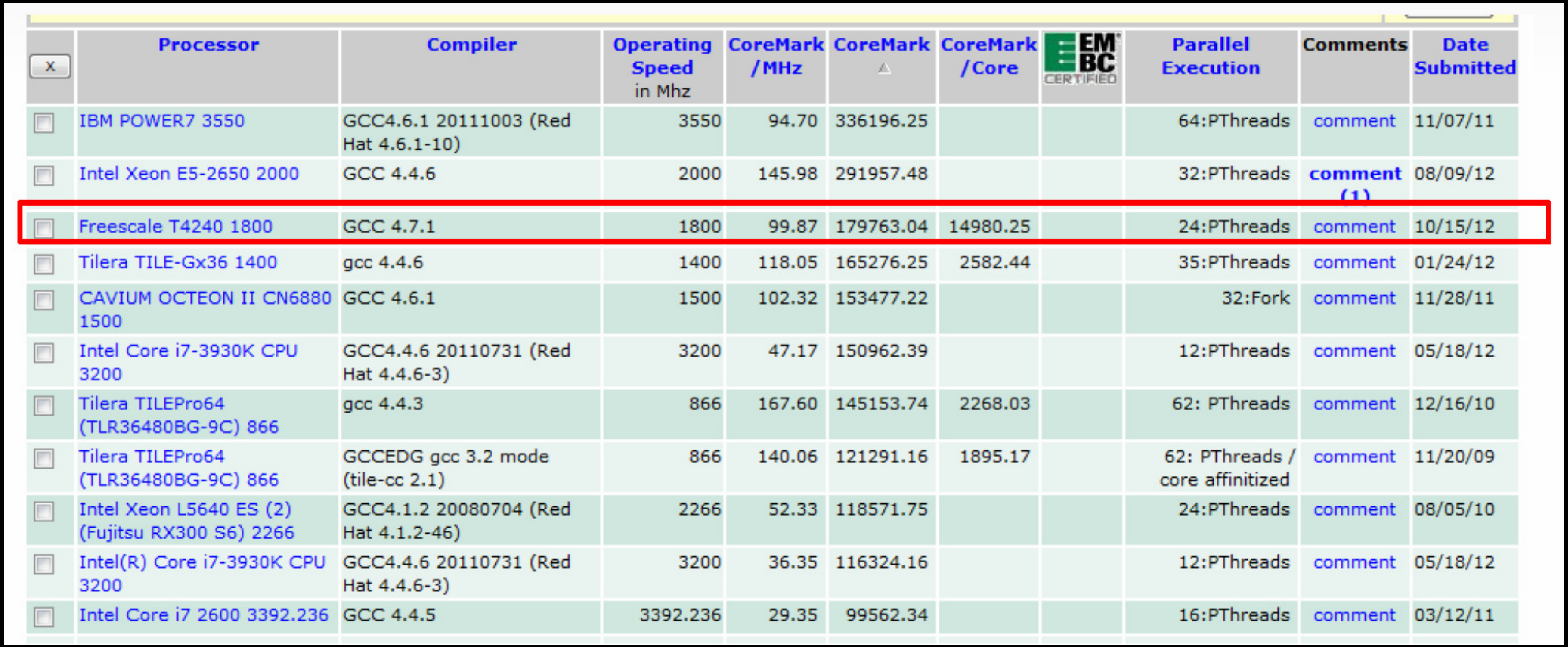
T4240 node board feasibility completed

T4240 node board planned 4Q2013

19” drawer planned 1Q14

PS. P5020 micro-web-server can be viewed here: http://www.swissdutch.ch:6999/

Performance comparison



x	Processor	Compiler	Operating Speed in Mhz	CoreMark /MHz	CoreMark △	CoreMark /Core	EEMBC CERTIFIED	Parallel Execution	Comments	Date Submitted
□	IBM POWER7 3550	GCC4.6.1 20111003 (Red Hat 4.6.1-10)	3550	94.70	336196.25			64:PThreads	comment	11/07/11
□	Intel Xeon E5-2650 2000	GCC 4.4.6	2000	145.98	291957.48			32:PThreads	**comment (1)**	08/09/12
□	Freescale T4240 1800	GCC 4.7.1	1800	99.87	179763.04	14980.25		24:PThreads	comment	10/15/12
□	Tilera TILE-Gx36 1400	gcc 4.4.6	1400	118.05	165276.25	2582.44		35:PThreads	comment	01/24/12
□	CAVIUM OCTEON II CN6880 1500	GCC 4.6.1	1500	102.32	153477.22			32:Fork	comment	11/28/11
□	Intel Core i7-3930K CPU 3200	GCC4.4.6 20110731 (Red Hat 4.4.6-3)	3200	47.17	150962.39			12:PThreads	comment	05/18/12
□	Tilera TILEPro64 (TLR36480BG-9C) 866	gcc 4.4.3	866	167.60	145153.74	2268.03		62: PThreads	comment	12/16/10
□	Tilera TILEPro64 (TLR36480BG-9C) 866	GCCEDG gcc 3.2 mode (tile-cc 2.1)	866	140.06	121291.16	1895.17		62: PThreads / core affinitized	comment	11/20/09
□	Intel Xeon L5640 ES (2) (Fujitsu RX300 S6) 2266	GCC4.1.2 20080704 (Red Hat 4.1.2-46)	2266	52.33	118571.75			24:PThreads	comment	08/05/10
□	Intel(R) Core i7-3930K CPU 3200	GCC4.4.6 20110731 (Red Hat 4.4.6-3)	3200	36.35	116324.16			12:PThreads	comment	05/18/12
□	Intel Core i7 2600 3392.236	GCC 4.4.5	3392.236	29.35	99562.34			16:PThreads	comment	03/12/11

Acknowledgements

This work is the results of many *people*

- Peter v. Ackeren, FSL
- Yvonne Chan, IBM Toronto
- Andreas Doering, IBM ZRL
- Tom Wilson, IBM Armonk
- Alessandro Curioni, IBM ZRL
- Stephan Paredes, IBM ZRL
- James Nigel, FSL
- Gary Streber, FSL
- Patricia Sagmeister, IBM ZRL
- Boris Bialek, IBM Toronto
- Marco de Vos, Astron NL
- Hillery Hunter, IBM WRL
- Vipin Patel, IBM Fishkill
- And many more remain unnamed….

Companies: FSL Austin, Belgium & China; IBM worldwide; Dsgnworx - NL