
# Prototyping in PRACE
# 5. PRACE Executive Industrial Seminar

Torsten Wilde (LRZ), 1IP-WP9 & 2IP-WP11 Leader, Bad Boll, Apr-2013

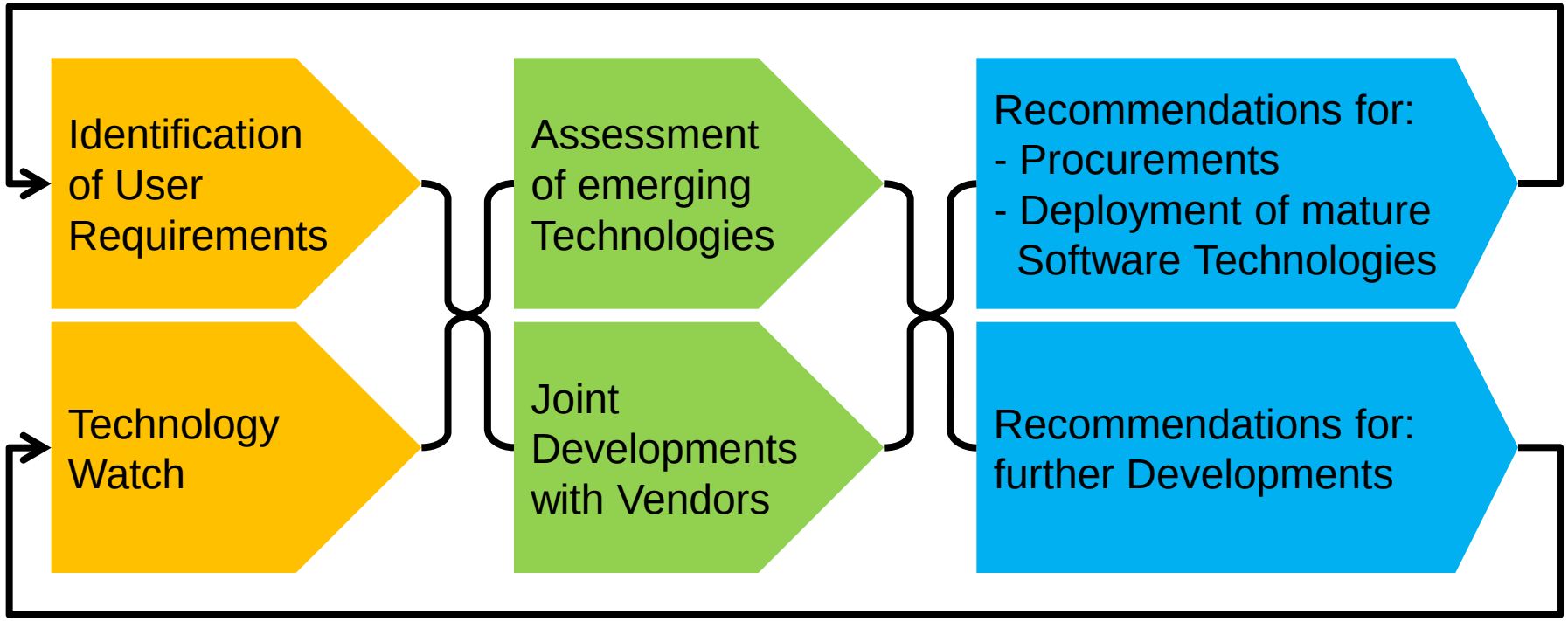
CAPACITIES

e-infrastructure

# Outline

- Prototyping in PRACE
- WP Objectives
- Results
  - T9.1 Technology Watch and Assessment
  - T9.2 Multi-Petascale Software Environment
  - T9.3 Future Highly energy efficient Tier-0 systems
- Conclusions

# Prototyping is Mandatory for PRACE

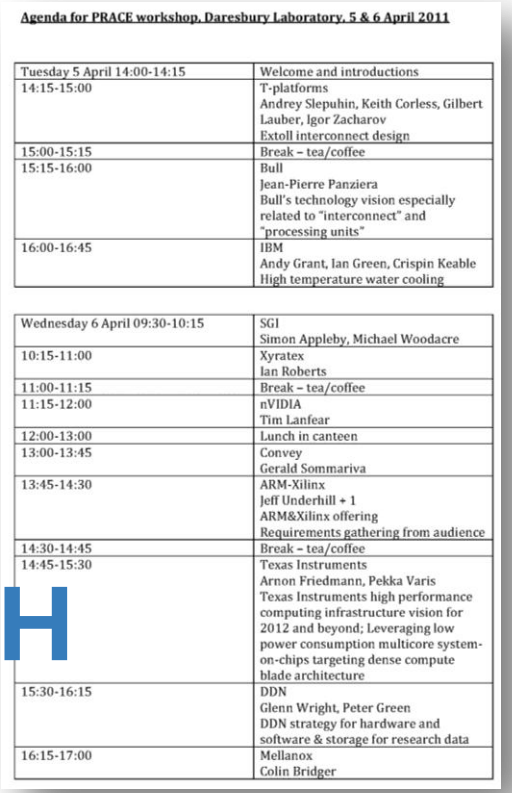


# Objectives of WP9

- Assessment of emerging hardware and software technologies for multi-Petascale systems (T9.1)
- Evaluation of future system software, programming paradigms and tools (T9.2)
- Evaluation of energy efficient HPC solutions (T9.3)

- **http://www.prace-ri.eu/Public-Deliverables?lang=en#PD_1IP**

# T9.1 TECHNOLOGY WATCH AND ASSESSMENT



**Agenda for PRACE workshop, Daresbury Laboratory, 5 & 6 April 2011**

| | |
|---|---|
| Tuesday 5 April 14:00-14:15 | Welcome and introductions |
| 14:15-15:00 | T-platforms<br>Andrey Slepuhin, Keith Corless, Gilbert Lauber, Igor Zacharov<br>Extoll interconnect design |
| 15:00-15:15 | Break – tea/coffee |
| 15:15-16:00 | Bull<br>Jean-Pierre Panziera<br>Bull's technology vision especially related to "interconnect" and "processing units" |
| 16:00-16:45 | IBM<br>Andy Grant, Ian Green, Crispin Keable<br>High temperature water cooling |

| | |
|---|---|
| Wednesday 6 April 09:30-10:15 | SGI<br>Simon Appleby, Michael Woodacre |
| 10:15-11:00 | Xyratex<br>Ian Roberts |
| 11:00-11:15 | Break – tea/coffee |
| 11:15-12:00 | nVIDIA<br>Tim Lanfear |
| 12:00-13:00 | Lunch in canteen |
| 13:00-13:45 | Convey<br>Gerald Sommariva |
| 13:45-14:30 | ARM-Xilinx<br>Jeff Underhill + 1<br>ARM&Xilinx offering<br>Requirements gathering from audience |
| 14:30-14:45 | Break – tea/coffee |
| 14:45-15:30 | Texas Instruments<br>Arnon Friedmann, Pekka Varis<br>Texas Instruments high performance computing infrastructure vision for 2012 and beyond; Leveraging low power consumption multicore system-on-chips targeting dense compute blade architecture |
| 15:30-16:15 | DDN<br>Glenn Wright, Peter Green<br>DDN strategy for hardware and software & storage for research data |
| 16:15-17:00 | Mellanox<br>Colin Bridger |

# Exascale Technology Assessment Results D9.1.2

1. Data and memory hierarchy – 3D stacking of memory, silicon photonics, deepening memory hierarchies
2. Fault tolerance – possible, at a price of increased hardware and running cost
3. Energy efficiency – the need for software optimisation
4. Architecture – heterogeneity, type of cores, interconnect
5. Scale – number of cores and massive parallelism

Consequence of future hardware technologies on software, assessment of programming languages and the system software

# T9.2 MULTI-PETASCALE SOFTWARE ENVIRONMENT

# Final Software Evaluation Report (D9.2.2)

- Three main aspects
  - *Programming Languages*
    for CPUs, GPUs and hybrid use.
  - *System Software*
    system and data management tools, MPI and other communication libraries and resource management.
  - *Computer Architectures*
    Recommendations and lessons learned from the prototype owners.

# Programming Languages for hybrid programming

- pragma-based: OpenACC / OmpSs
  - Boost productivity, not ready for production level software
  - #pragma acc kernels copyin(temp_in[0:ni*nj]) copy(temp_out[0:ni*nj])
- OpenCL / CUDA:
  - OpenCL closed performance gap but is lacking support for latest hardware features
- UPC (Unified Parallel C):
  - low level knowledge is required in order to write performant code

# Tools

- Tools for hybrid CPU/GPU systems
  - Support hybrid application debugging
  - Memory tracing
  - Profiling
- I/O tools
  - I/O will be bottleneck at Exascale, need profiling tools for I/O operations
- Usability of future tools
  - Exponential increase of available information at Exascale requires a re-thinking of data presentation
    - Filtering traces from millions of running threads
    - Representing results

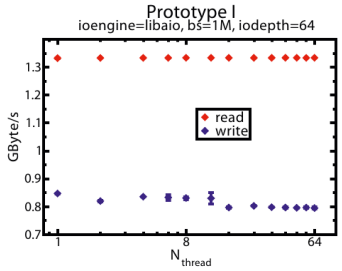
Energy benchmark, prototype assessment and status

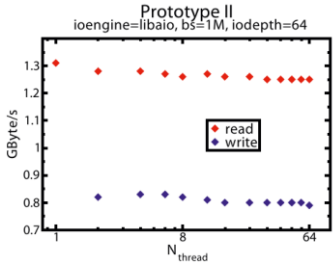
# T9.3 FUTURE HIGHLY ENERGY EFFICIENT TIER-0 SYSTEMS

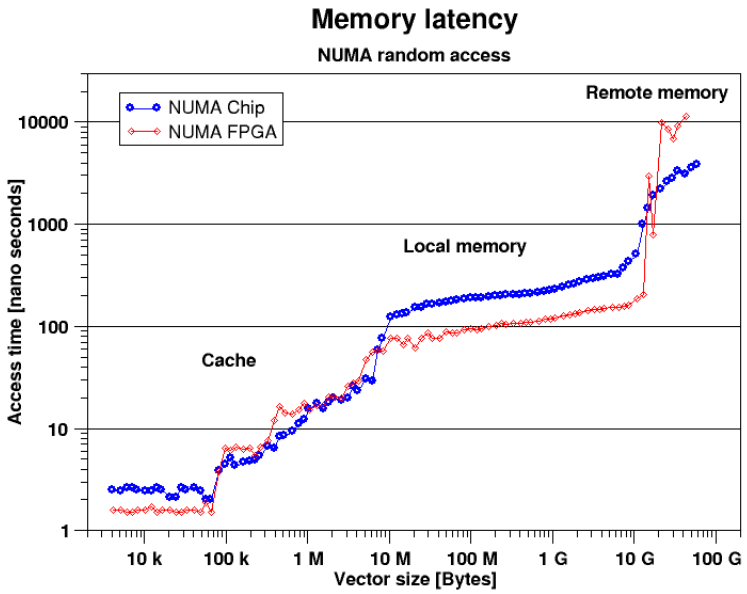
# 1st Set of Prototypes

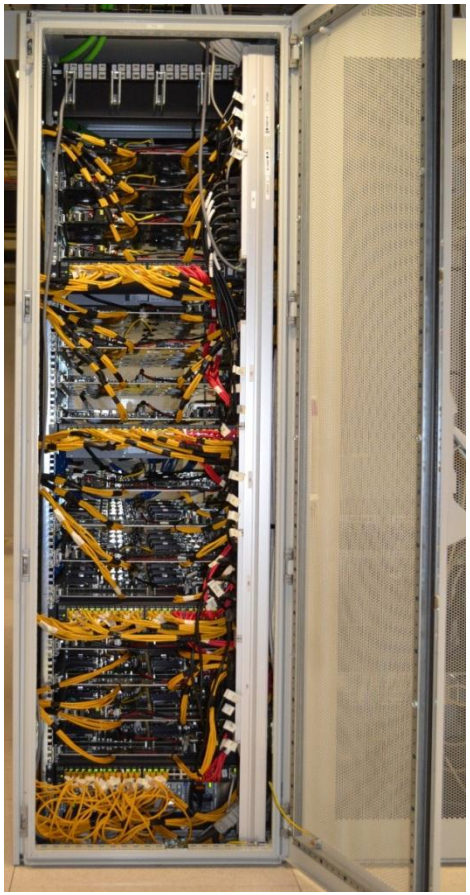
Neo 3000 Xyratex prototype at CEA

- Memory and I/O
  - Exascale I/O and MPP I/O Prototypes up and running
    - Working on common benchmarks
  - NUMA-CIC
    - 1st phase running, waiting for final NUMA hardware revision
- Energy To Solution
  - ARM, LRZ – Hot Water Cooled and SNIC-DSP Prototypes up and running
    - Writing final reports till end of October
- Accelerators:
  - GPGPU and Hybrid CPU/GPU up and running
    - GPGPU will provide final report by end of September
    - CPU/GPU is evaluating the 2nd phase hardware

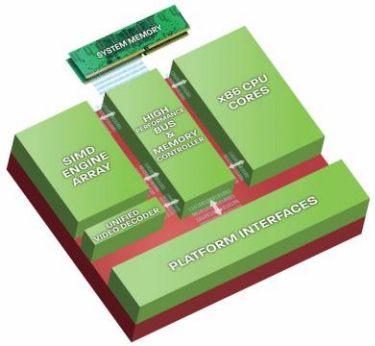


# 1ST SET OF PROTOTYPES RESULTS



# Node technologies

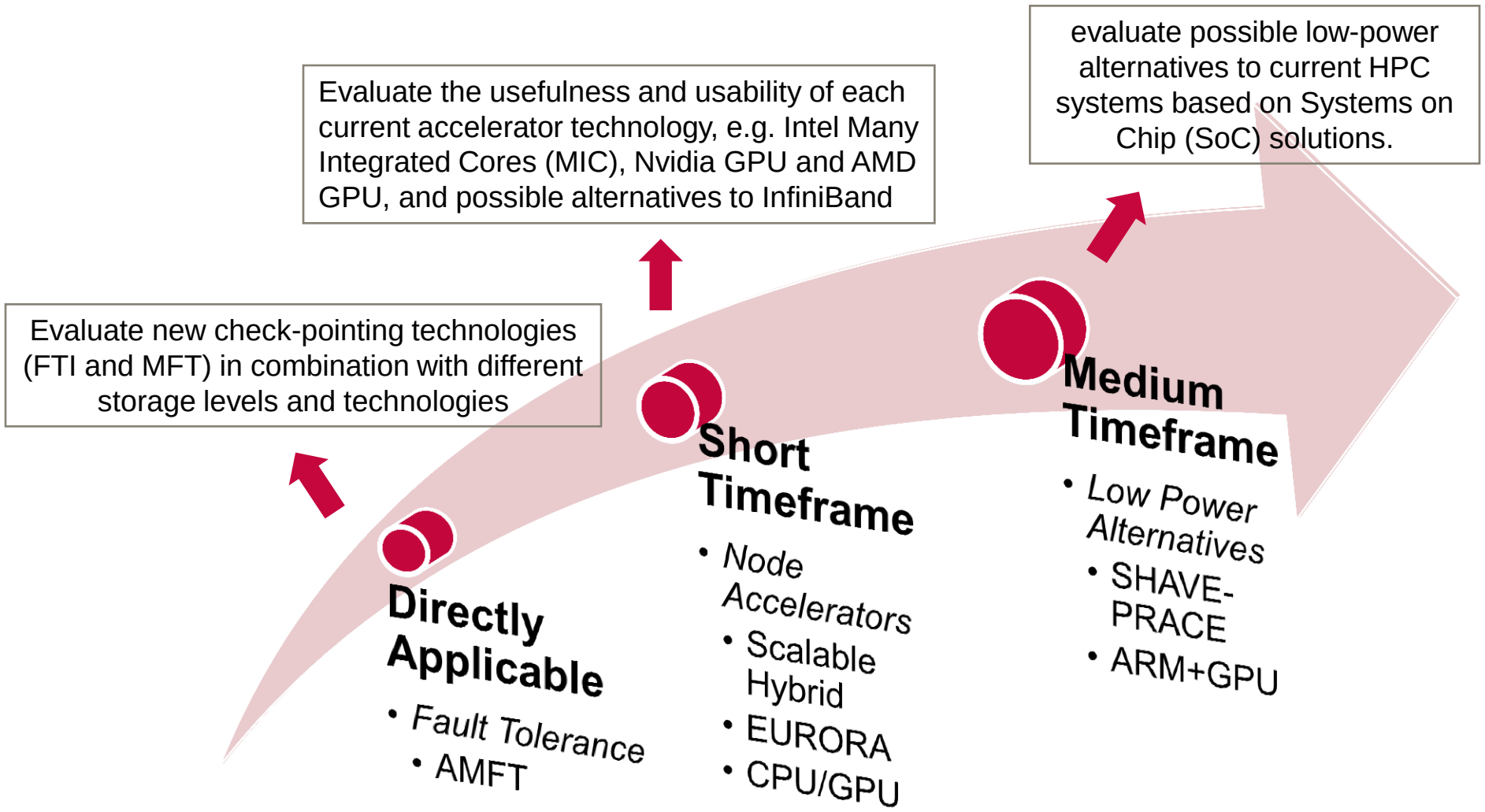
- optimized implementation of  matrix-multiplication
  - FPGA: 5 – 10 times higher energy efficiency than an x86 software solution
  - DSP: 2 – 5 times
  - Programming effort very high
- First and second generation x86+GPU APUs
  - not competitive in regards to  energy efficiency
  - Benchmarking behavior similar to GPUs
- A HPC cluster can be build with mobile SoC (ARM)
  - floating-point capabilities of the current generation mobile CPUs is insufficient to be competitive in terms of energy efficiency

# I/O

- tighter integration of I/O and file systems into MPP and cluster systems and the use of flash technology is beneficial
- The concept of using flash and disks in a tiered file system is working in GPFS.
- Good performance dependents on good software implementations
- There are efficiency and scalability issues related to large file systems in which a large fraction of the files are small.

# Benchmarking results

- direct relationship between performance and energy efficiency for most benchmarks
  - Software optimization most important
- highest performing hardware platform does not, in general, minimize the energy consumption for a task
- Direct hot water cooling is viable
  - Re-use via adsorption possible


evaluate possible low-power alternatives to current HPC systems based on Systems on Chip (SoC) solutions.
Evaluate the usefulness and usability of each current accelerator technology, e.g. Intel Many Integrated Cores (MIC), Nvidia GPU and AMD GPU, and possible alternatives to InfiniBand
Evaluate new check-pointing technologies (FTI and MFT) in combination with different storage levels and technologies
Directly Applicable
• Fault Tolerance
• AMFT
Short Timeframe
• Node Accelerators
• Scalable Hybrid
• EURORA
• CPU/GPU
Medium Timeframe
• Low Power Alternatives
• SHAVE-PRACE
• ARM+GPU

# Conclusion

- Future technology:
  - Flash based I/O devices
  - Warm water cooling
  - Prove of concept, low-power commodity parts can be used for HPC
- New technologies are important for PRACE
- Results guide development of PRACE-RI