Addressing the Challenge of Exascale

European Exascale Software Initiative EESI

Towards Exascale roadmap implementation

EESI2 – Recommendations
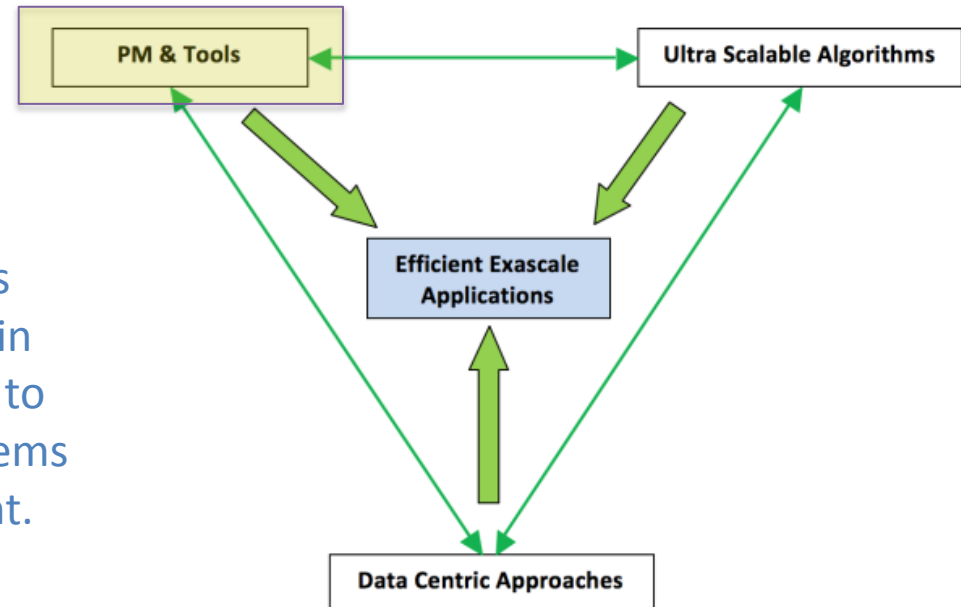
Holistic Approach to Resiliency

Ferad Zyulkyarov, EESI2 WP5

Barcelona Supercomputing Centre

EESI2 recommendations

New holistic approach to resiliency which spans across all software layers and abstracts the heterogeneity within and between the systems is required to keep future **Extreme computing** systems operable and **Extreme data** consistent.

New **disruptive** resiliency techniques based on emerging technologies and innovations should be developed that deliver **several orders of magnitude more efficiency** and **increased productivity** in building resilient tools and applications for Exascale.
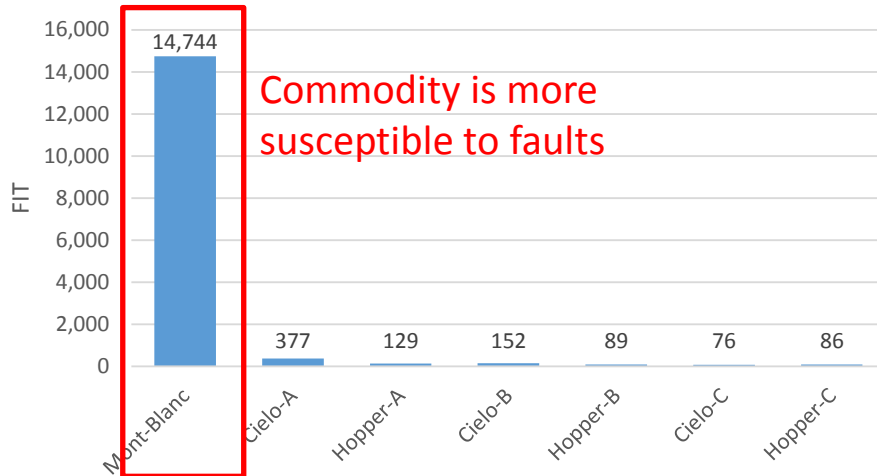
Toward more reliable Extreme Computing
with holistic approach to resiliency

Motivations

❖ Future Exascale systems are expected to see a **fault in sub second intervals**
  - Today's Petascale systems are subject one fault every 35 min (DARPA, Kogge 2008)

❖ Most of the down time is due to errors in the software
  - 53% of the down in Blue Waters is due to the SW (Di Martino et al DSN'2014)

❖ Existing resiliency techniques along the SW stack do not integrate well and sometimes are incompatible
  - RMA in MPI 3.0 is incompatible with message logging (ESSI2 Deliverable D5.1)

❖ Existing techniques may not scale to Exascale systems
  - Checkpointing Extreme Data would take too long

❖ Heterogeneity is another challenge for building reliable Exascale systems
  - System components do not report and provide uniform set of reliability features
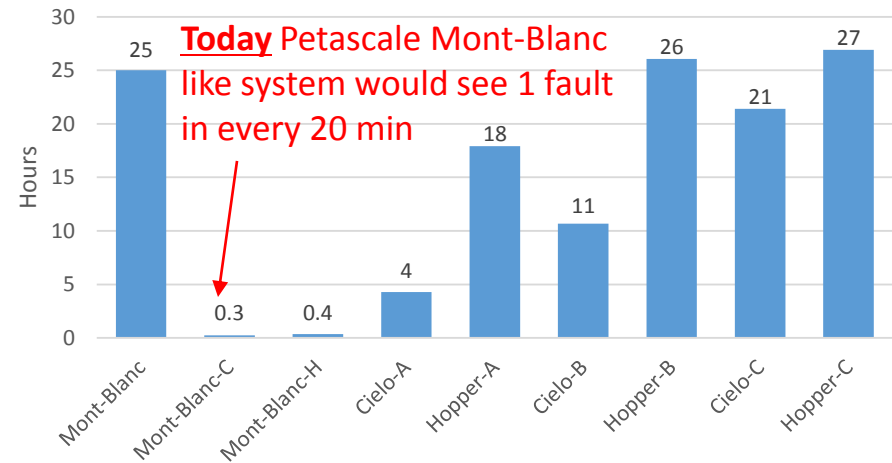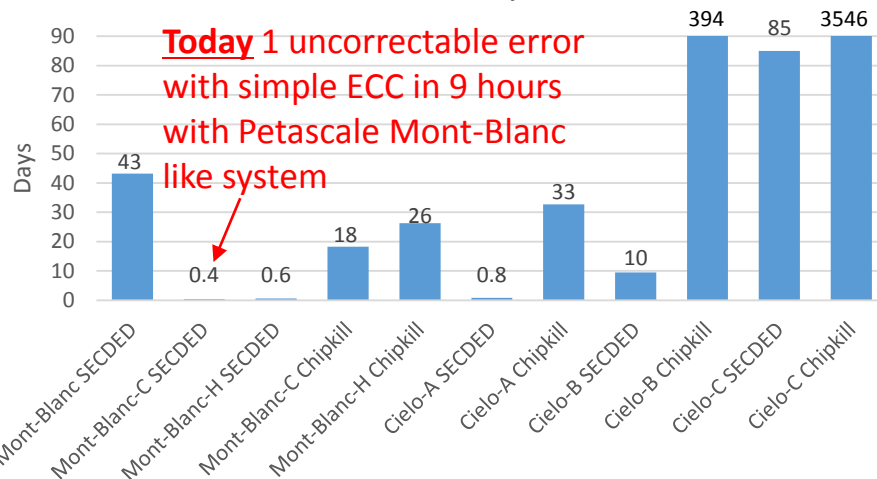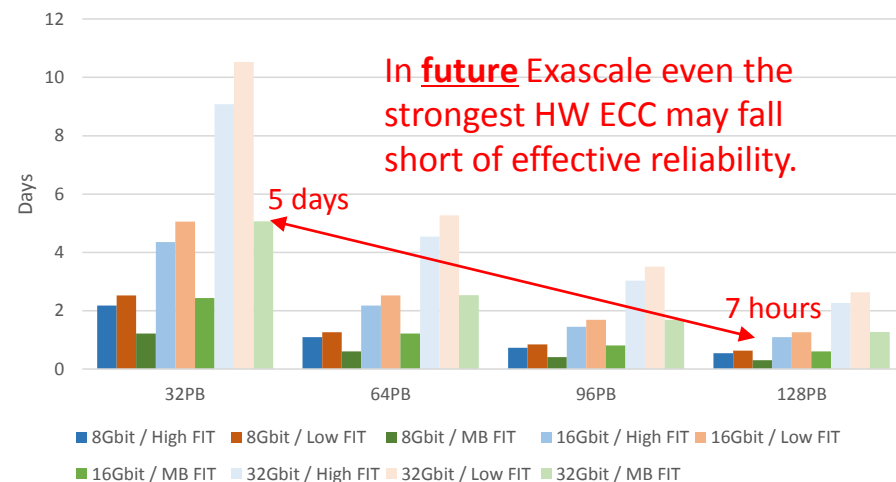
Examples: Memory reliability

Memory FIT for 1GB

Commodity is more susceptible to faults

14,744 (Mont-Blanc), 377 (Cielo-A), 129 (Hopper-A), 152 (Cielo-B), 89 (Hopper-B), 76 (Cielo-C), 86 (Hopper-C)

Memory faults (corrected & uncorrected) MTBF

Today Petascale Mont-Blanc like system would see 1 fault in every 20 min

25 (Mont-Blanc), 0.3 (Mont-Blanc-C), 0.4 (Mont-Blanc-H), 4 (Cielo-A), 18 (Hopper-A), 11 (Cielo-B), 26 (Hopper-B), 21 (Cielo-C), 27 (Hopper-C)

Uncorrectable Memory Faults MTBF

Today 1 uncorrectable error with simple ECC in 9 hours with Petascale Mont-Blanc like system

43 (Mont-Blanc SECDED), 0.4 (Mont-Blanc-C SECDED), 0.6 (Mont-Blanc-H SECDED), 18 (Mont-Blanc-C Chipkill), 26 (Mont-Blanc-H Chipkill), 0.8 (Cielo-A SECDED), 33 (Cielo-A Chipkill), 10 (Cielo-B SECDED), 394 (Cielo-B Chipkill), 85 (Cielo-C SECDED), 3546 (Cielo-C Chipkill)

Exascale projection MTBF (High FIT is Vendor A)

In **future** Exascale even the strongest HW ECC may fall short of effective reliability.

5 days

7 hours

Legend: 8Gbit / High FIT, 8Gbit / Low FIT, 8Gbit / MB FIT, 16Gbit / High FIT, 16Gbit / Low FIT, 16Gbit / MB FIT, 32Gbit / High FIT, 32Gbit / Low FIT, 32Gbit / MB FIT

X-axis: 32PB, 64PB, 96PB, 128PB

And this is only for memory. What about CPUs, network, storage and software?

PM & Tools

Proposal : Fund R&D programs in order to explore

❖ Fault model in large heterogeneous systems for Extreme Computing

❖ Resiliency API to:

- integrate the different resiliency techniques across software stack

- provide uniform HW/SW interface to abstract the heterogeneity in hardware

❖ New disruptive checkpoint/restart techniques based on non-volatile memory technologies and new memory architectures

❖ Multi-level hierarchical fault-tolerance techniques

❖ New best practices for developing resilient software and tools for Exascale using new programming models and languages such as task-based parallel programs

❖ Effective and efficient failure prediction for Exascale that can run online

❖ Design and implementation of disruptive techniques for redundant execution such as selective replication that require only fraction of the system's resources

Importance and Emergency

- **Impact on the efficiency of future Exascale applications**
    - The recommendations will enable the development of reliable and fault-tolerant Extreme Computing and Extreme Data applications that scale to and beyond Exascale

- **Emergency for starting this work**
    - Work on the recommendations should start as soon as possible

- **Impact on improving EU strength**
    - This work will significantly improve EU's competitiveness and strengthen its position as a leader in HPC

R&D Efforts

- ❏ **Common scientific platform on resilience for Exascale**
  - ■ to align the research on resiliency and collaborate

- ❏ **Resources**
  - ■ 80 person years (20 people)

- ❏ **Funding**
  - ■ 8M €

- ❏ **Duration**
  - ■ 4 Years