Error Permissive Computing: a New Approach for Post Moore's Computer System Design

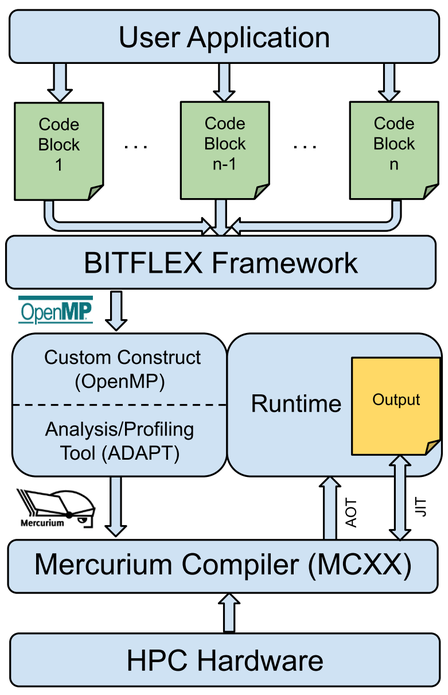
Ryousei Takano, Takahiro Hirofuchi, Mohamed Wahib,
Truong Thao Nguyen, Hiroki Kanezashi, Akram Ben Ahmed
National Institute of Advanced Industrial Science and Technology

Abstract We are exploring a new concept of error permissive computing that improves the capability and capacity while drastically reducing power consumption. More specifically, we controllably allow hardware errors and develop system software to assure acceptable computational results. For example, an error correction technique can result in increased latency and reduced capacity. By taking a holistic approach across the layers from hardware to software, lightweight and appropriate error correction is performed at the software layer while eliminating general purpose error correction in hardware layer.

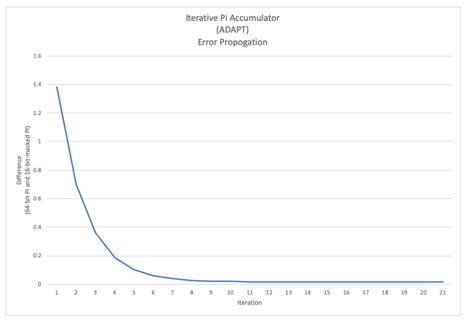
BITFLEX: A framework to enable error permissive computing [1]

- We require an attractive means of **boosting performance and maintaining accuracy** in non-deterministic applications.
- **Solution**: BITFLEX framework incorporated in MCXX compiler.
- We propose an extension of OpenMP as follows:
  #pragma omp nondeter <parameters>

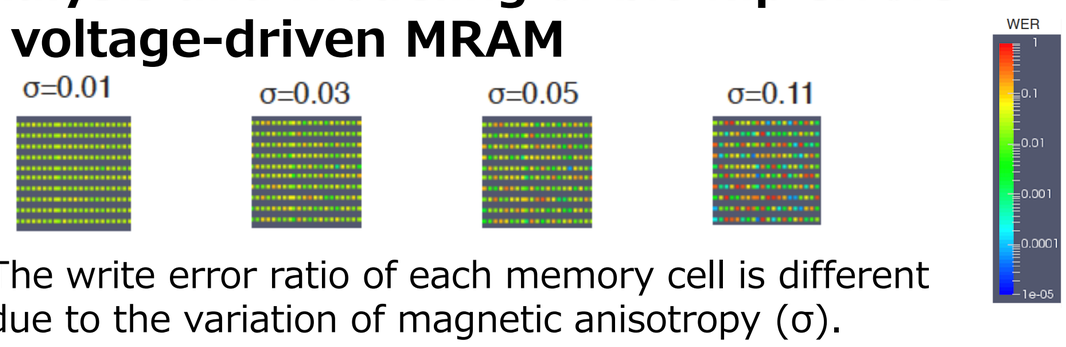
BITFLEX Full Stack
(OpenMP Extension)



ADAPT Case Study: Pi Accumulator



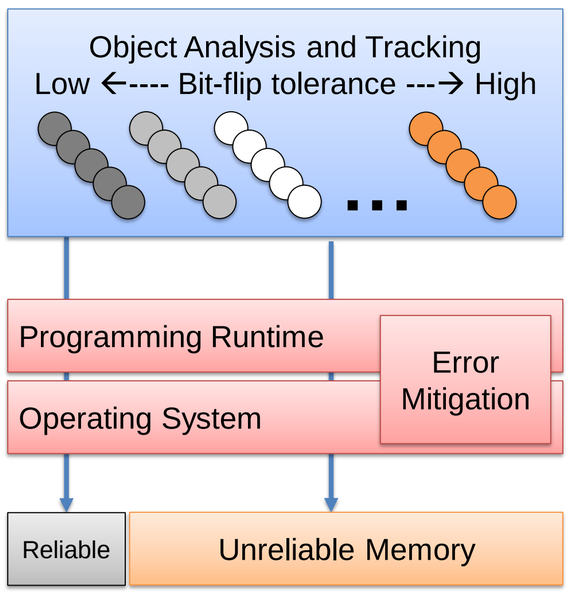
Analysis and modeling of bit-flip errors in voltage-driven MRAM

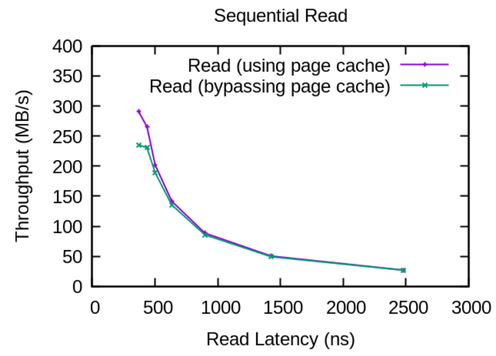


- The write error ratio of each memory cell is different due to the variation of magnetic anisotropy (σ).

FPGA-based new memory device emulator [2]

- Emulate the behavior of new memory devices (latency, bandwidth, bit error ratio) with high accurate.
- Enable detailed performance evaluation of new system software mechanisms.



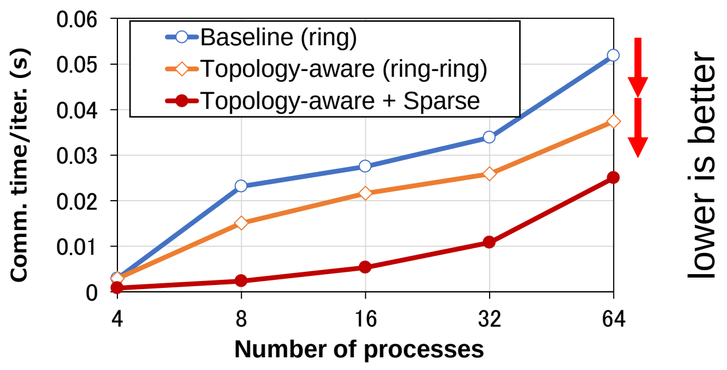


Accelerating communication for large-scaler deep learning [3]

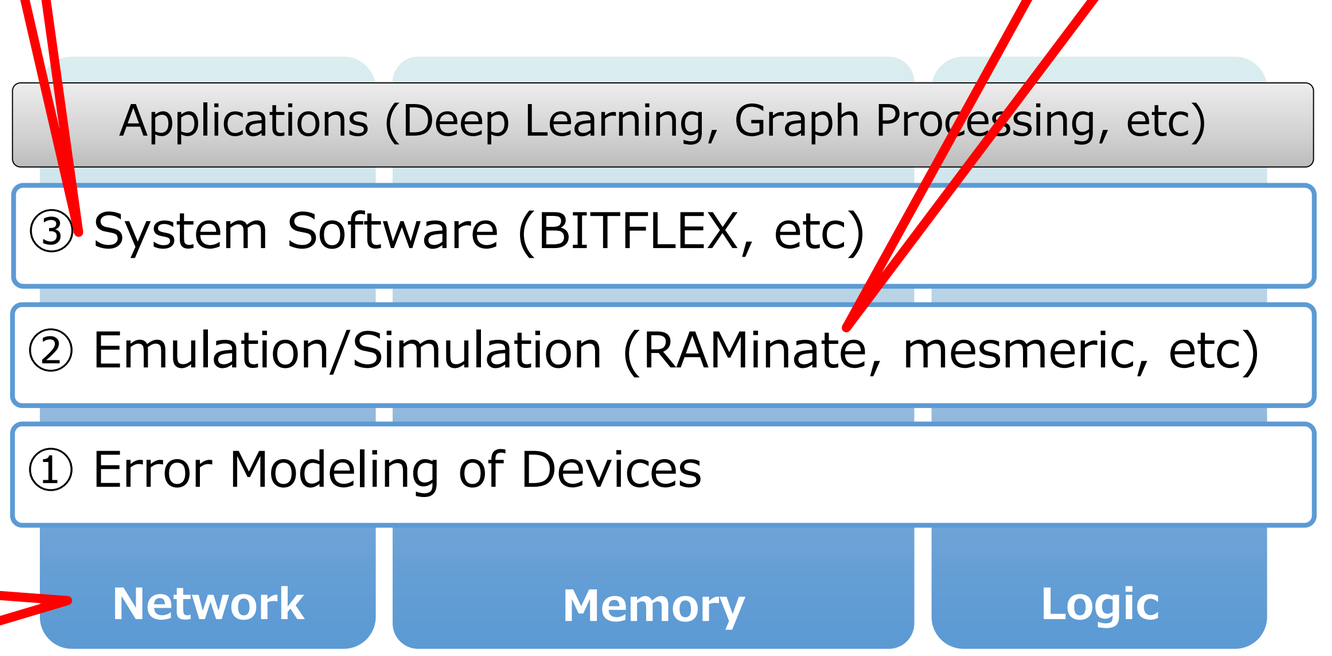
Topology-aware Allreduce
- ✓ Reduce comm. time up to 45%
- ✓ Reduce power consumption of comm. up to 23%

×

Sparse communication
- ✓ 100x-1000x compressed
- ✓ Reduce communication time ~40% more



Simulated result with ABCI-system, 32MB-message, 0.78% sparcification



References
[1] R. Barton, et al. "BITFLEX: A Dynamic Runtime Library for Bit-Level Precision Manipulation and Approximate Computing," HPC Asia 2020.
[2] T. Hirofuchi, et al. "FPGAによる次世代メモリのエミュレーション機構の試作", IPSJ SIGHPC171, 2019.
[3] T. Nguyen, et al. "Topology-aware Sparse Allreduce for Large-scale Deep Learning", IEEE IPCCC 2019.