rebellions_

LightTrader : World-first AI-enabled High-Frequency Trading Solution & 16 TFLOPS / 64 TOPS Deep Learning Inference Accelerators

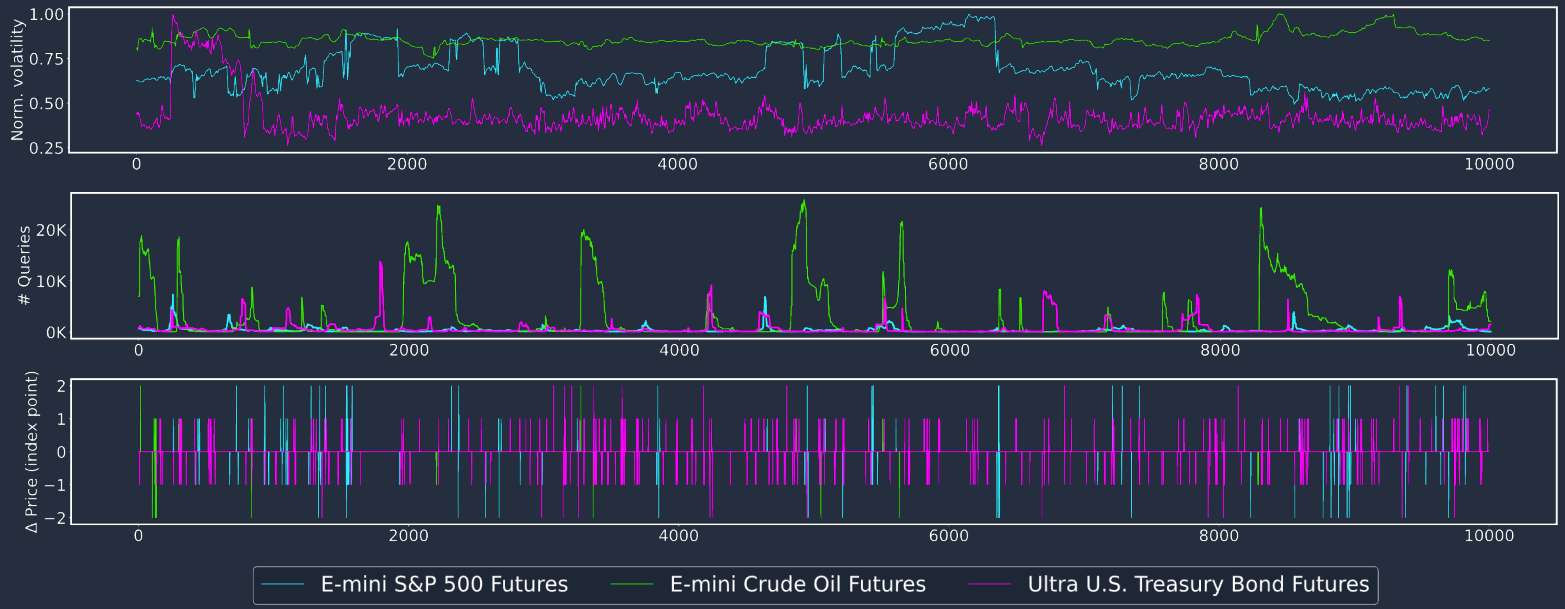
Hyunsung Kim, Sungyeob Yoo, Jaewan Bae, Kyeongryeol Bong, Yoonho Boo, Karim Charfi, Hyo-Eun Kim, Hyun Suk Kim, Byungjae Lee, Jaehwan Lee, Sungho Shin, Joo-Young Kim, Sunghyun Park, Jinwook Oh

World-first AI-enabled High Frequency Trading (HFT) Solution

We present the word-first AI-enabled HFT system, LightTrader, which integrates the custom AI accelerators and the FPGA-based conventional HFT pipeline for the short-latency-high-throughput trading solutions with a reduced query miss rate. For better utilization, adaptive job scheduling methods are also proposed to further improve the performance, where layer-wise workload scaling and dynamic voltage-frequency scaling (DVFS) techniques progressively adjust the workloads of AI accelerators, in conjunction with the architecture support.

Using historical data from Chicago Mercantile Exchange (CME) as a validation test set, LightTrader integrating TSMC 7nm tape-out accelerators solely achieves 6x speed-up of DNN processing and 30-50x reduction of query miss rate without the scheduling method while the scheduling scheme further improves the energy efficiency by 25% and reduces the query miss rate by 2.4x.

HFT Craves Deep Learning with Breakthrough Technologies of Reducing Latency & Miss Rate, in the Extremely Strict Constraints



Complicated market data patterns

Unpredictable profit opportunities

Burst of input query for AI processing

LightTrader makes it feasible to exploit the superb performance of DNN algorithms in HFT strategies, which have been impossible for conventional systems

Heterogeneous HFT System on Board : FPGA + AI Accelerators

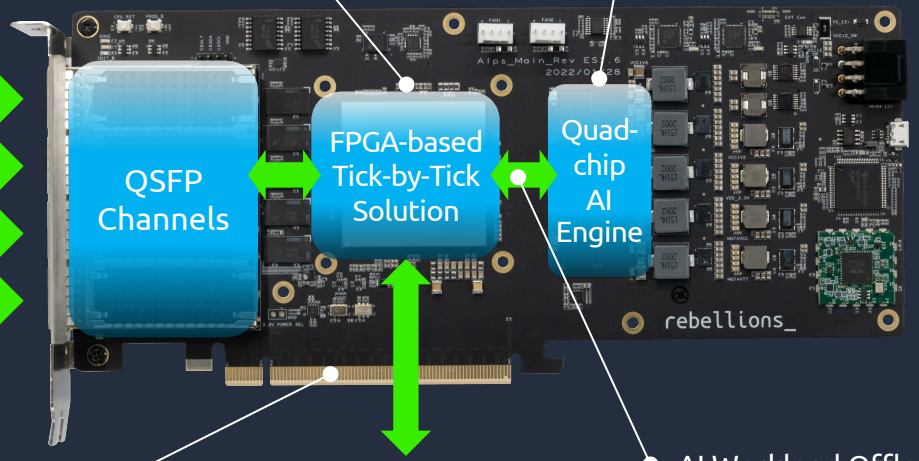
Conventional HFT Pipeline +
AI Offloading Engine
Intelligent Batch Selection
Layer-wise Workload Scaling

HFT-optimized AI Accelerators
Multi-Chip Intergration
Batch-insensitive Latency
Flexible AI Core Dataflow

Exchange

Multiple Market Data Feeds

Host Communication



Power Management Strategy
Colo Server Power Compatible DVFS

PCIe HFT Device Driver

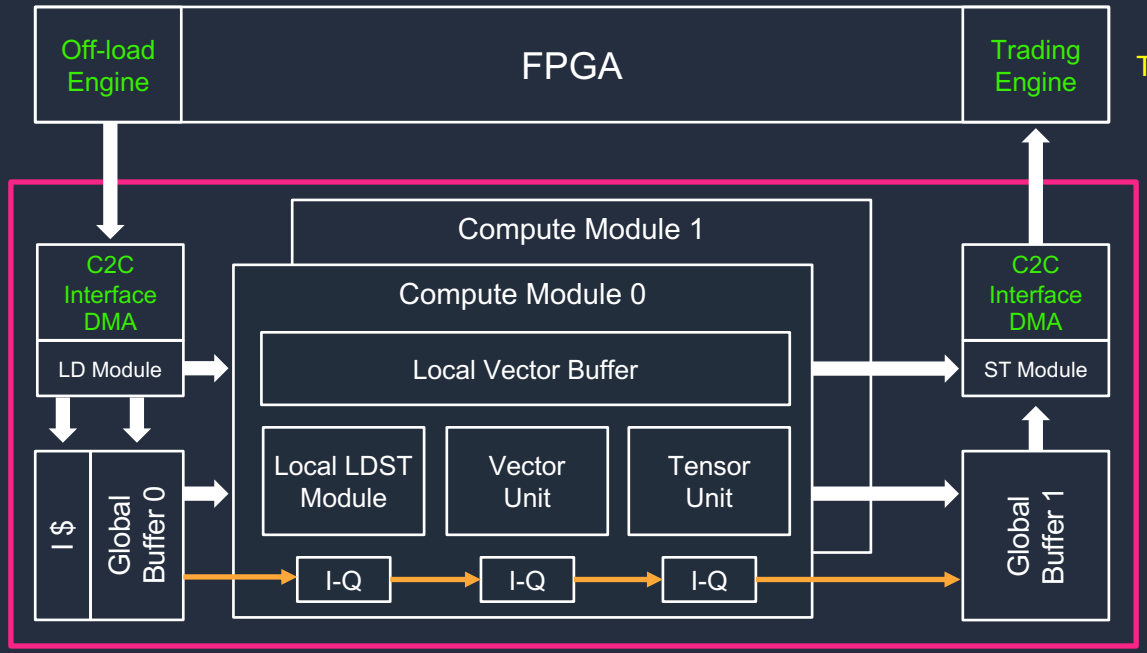
AI Workload Offloading
On-the-fly Feature Map Gen & Transfer
Highly Accurate Price Prediction

Energy-efficient Companion AI Accelerators with Dynamic Workload Scaling

On-the-fly Feature Map Gen

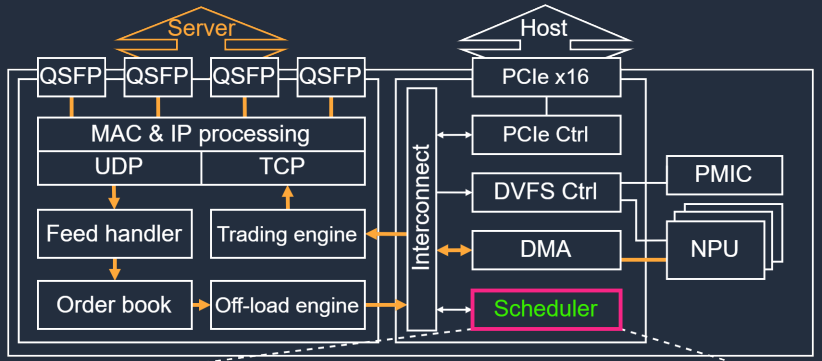
Custom C2C Protocol

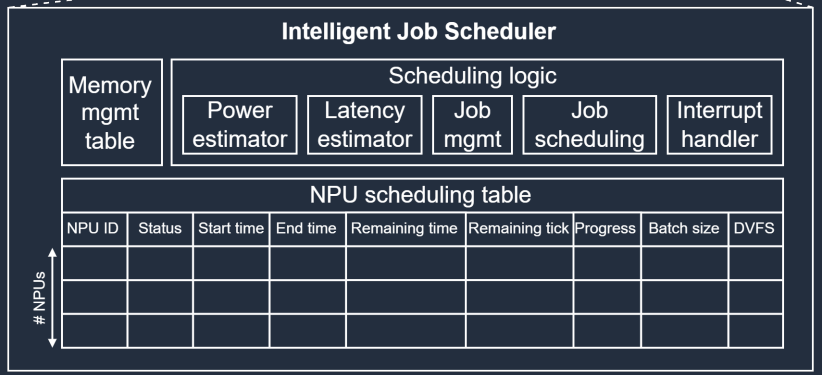
Dynamic Workload Control



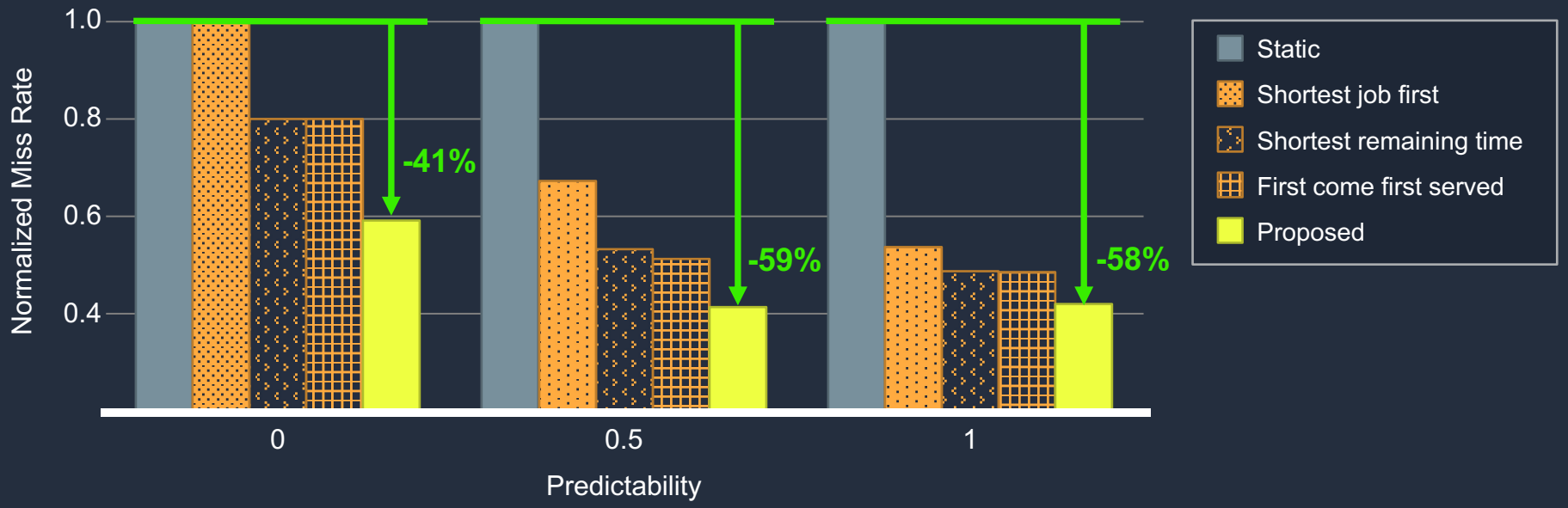
Flexible Dataflow/Precision (FP16, INT8) Computation

Prediction Horizon aware AI Task Scheduler



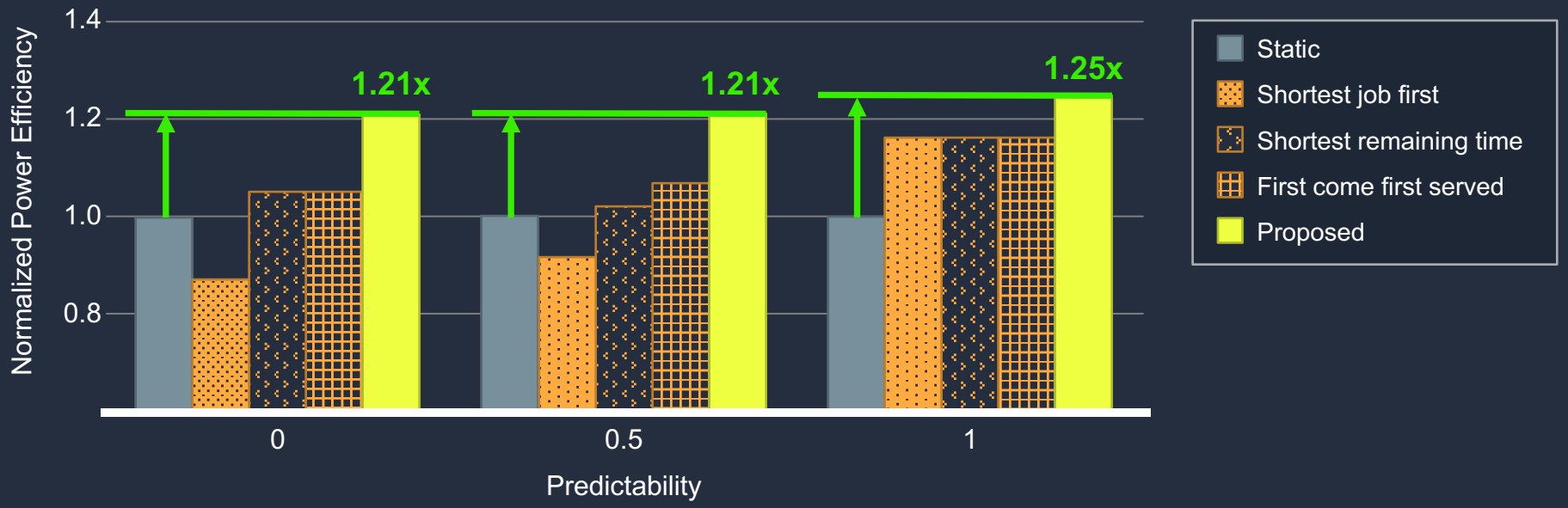


Miss Rate Reduction from the Intelligent Job Scheduling & DVFS



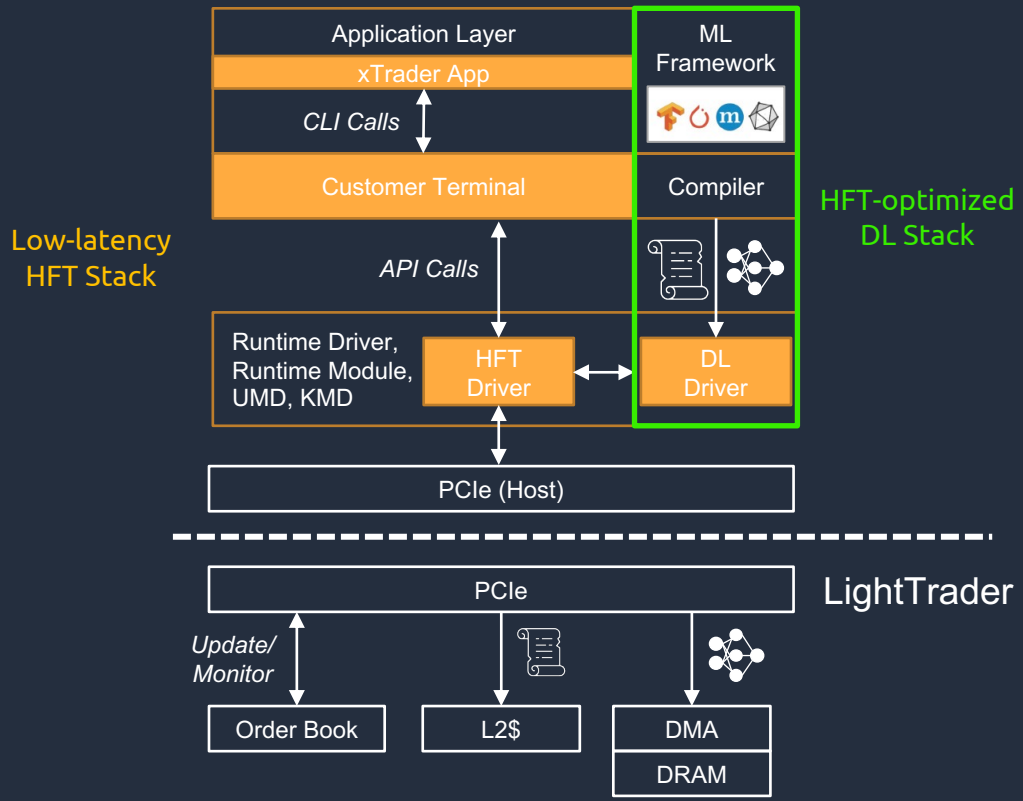
The reduced miss rate results in the higher profit

Efficient Power Control for Higher Throughput



The more efficient power utilization strategy achieves the higher profit

AI-enabled HFT Software Stack with RebelFlow



The HFT software stack provides an end-to-end solution to enable the low-latency finance AI model processing on the LightTrader hardware

The software stack includes
- Compiler
- DL Inference driver
- HFT pipeline driver
- User application interface

The PCIe Gen 4.0 x16 interface enables a real-time hardware control and monitor

Prototype System Integration for AI-based HFT Compute Node

Integrating eight board into a standard 4U rack form-factor, the proposed server-level solution extends the capability of the LightTraders up to

- 128 TFLOPS / 512 TOPS
- 3.2 Tbps query processing throughput
- 10~100 us DL inference-based tick-to-trade latency
- only with 35x8 W card power

Thank You