rebellions_

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

Hyunsung Kim[1]*, Sungyeob Yoo[2]*, Jaewan Bae[1], Kyeongryeol Bong[1], Yoonho Boo[1], Karim Charfi[1], Hyo-Eun Kim[1], Hyun Suk Kim[1], Jinseok Kim[1], Byungjae Lee[1], Jaehwan Lee[1], Myeongbo Shim[1], Sungho Shin[1], Jeong Seok Woo[1], Joo-Young Kim[2], Sunghyun Park[1], Jinwook Oh[1]

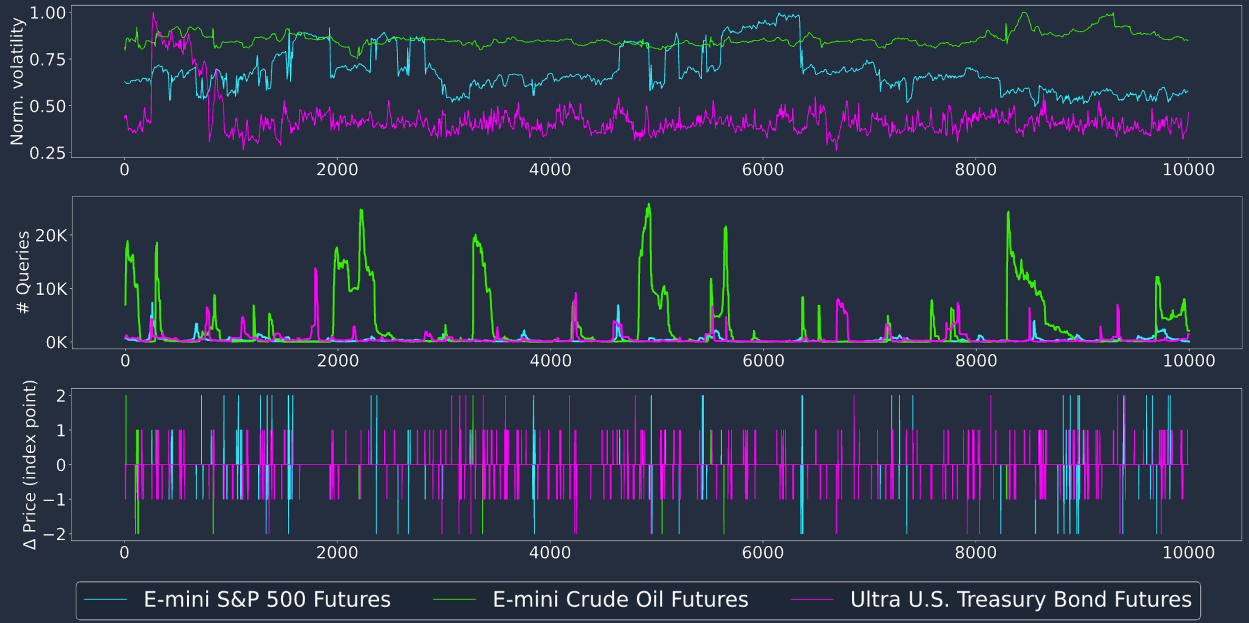
[1] Rebellions Inc., [2] KAIST
*These authors contributed equally to this work

HOT CHIPS

# 1-Pager Abstract

We present the world's first AI-enabled high-frequency trading (HFT) system, LightTrader, which integrates the custom AI accelerators and the FPGA-based conventional HFT pipeline for the low-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. 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 leveraging AI algorithms with breakthrough technologies of realizing low-latency-high-throughput, under the extremely strict constraints



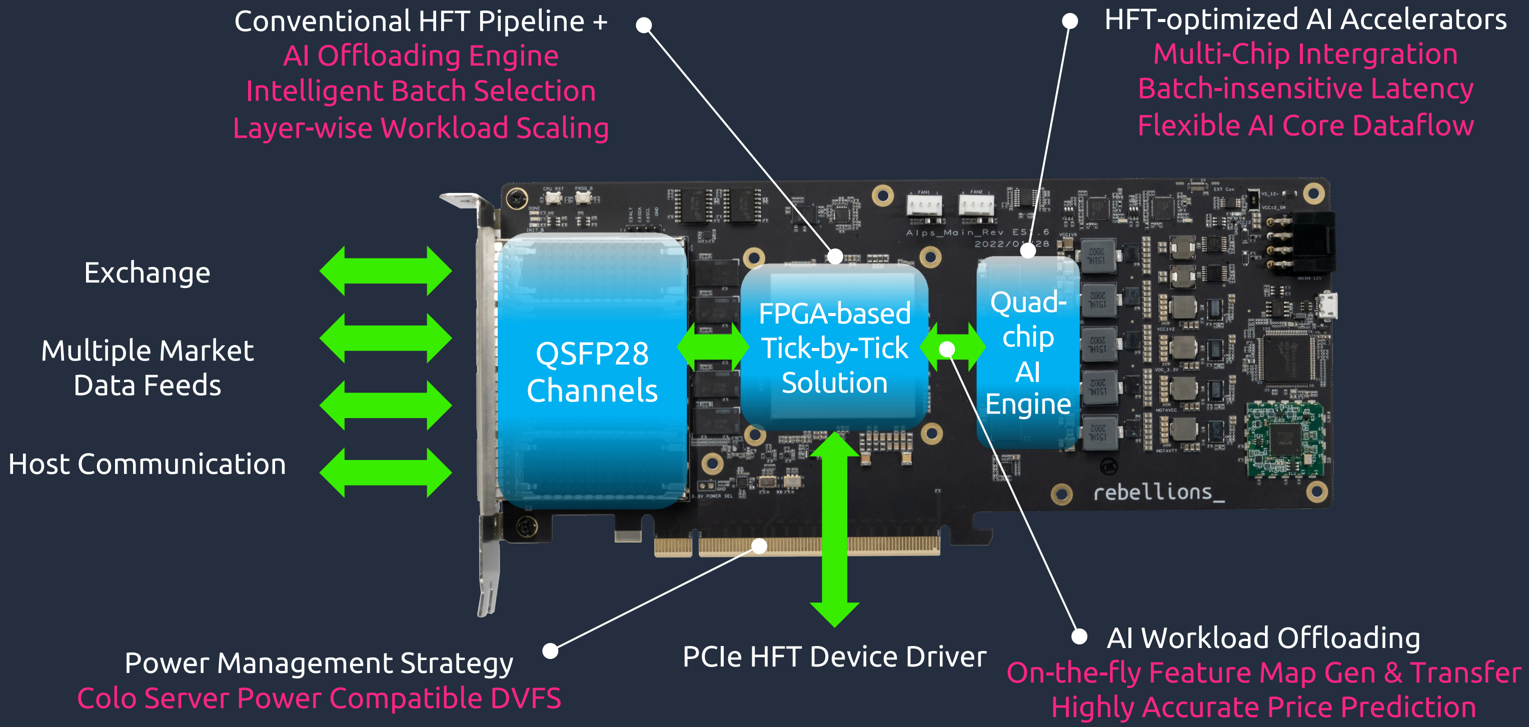
Complicated market data patterns

Bursty input query for AI processing

Unpredictable profit opportunities

LightTrader can exploit the superb performance of AI for HFT strategies, which have been impossible for conventional systems

# Heterogeneous HFT system on a 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

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

# Low-latency companion AI accelerators : 4 TFLOPS / 16 TOPS from single chip

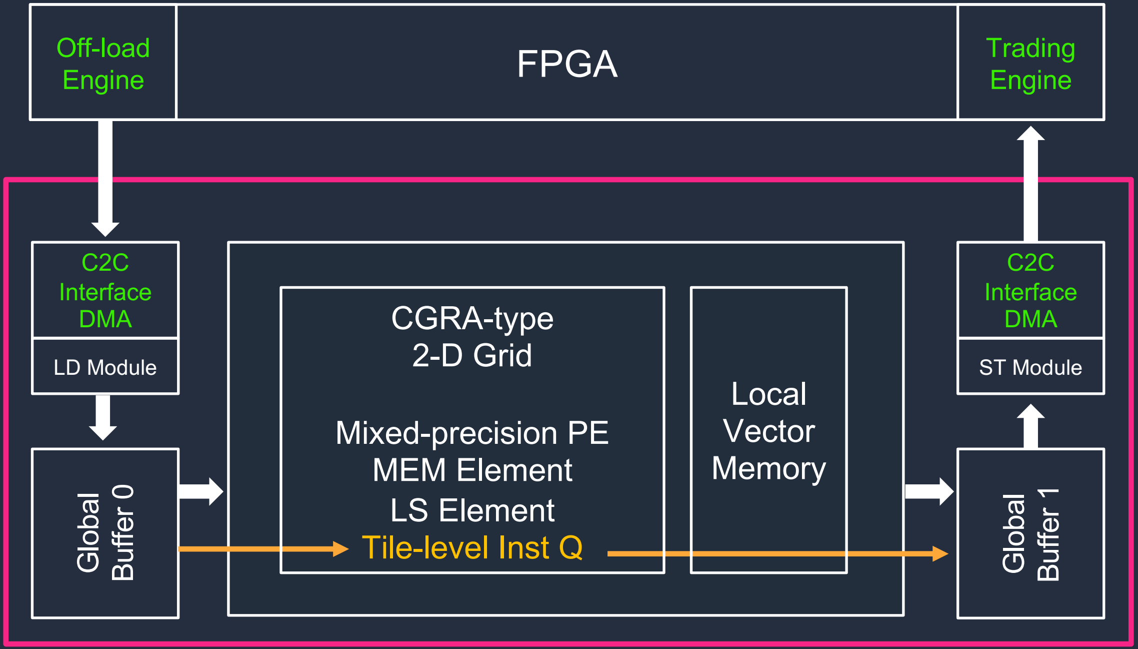
On-the-fly Feature Map Gen

Custom C2C Protocol

Dynamic Workload Control

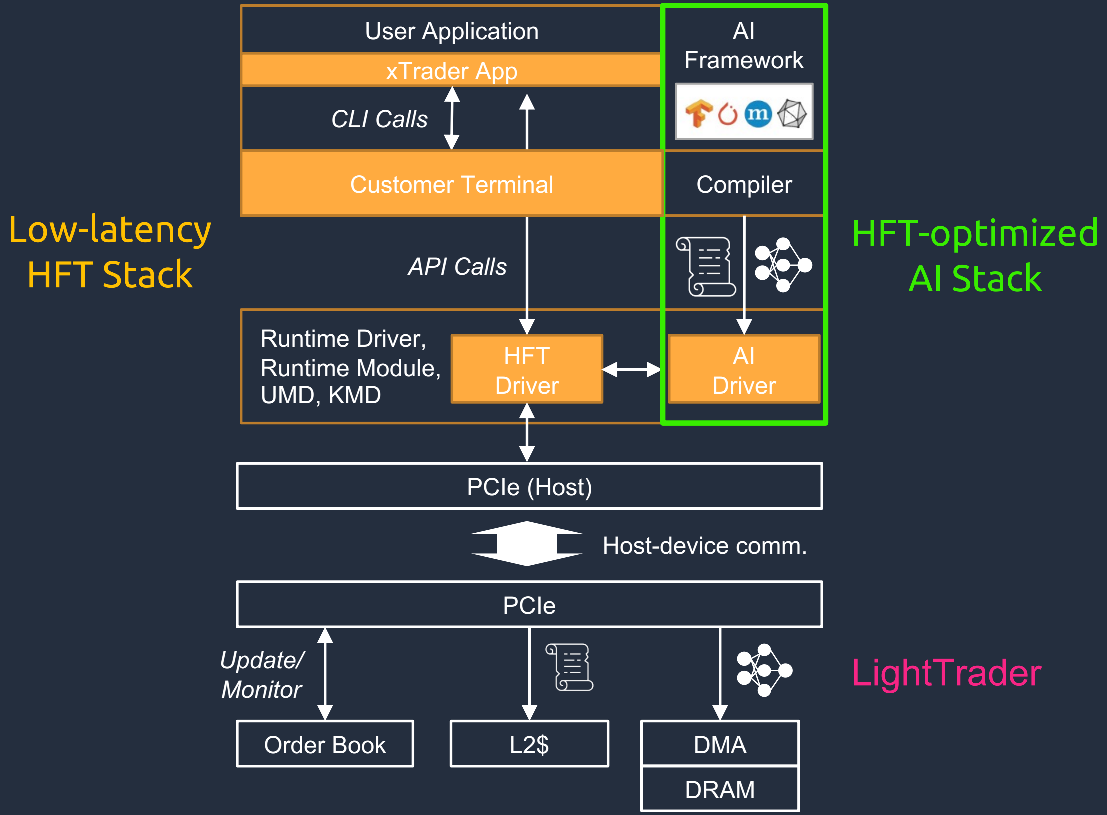
Trading Strategy

Inference Results (price, volatility, risk, etc.)



Flexible Dataflow/Precision (FP16, INT8) Computation

# AI-enabled HFT software stack with User API



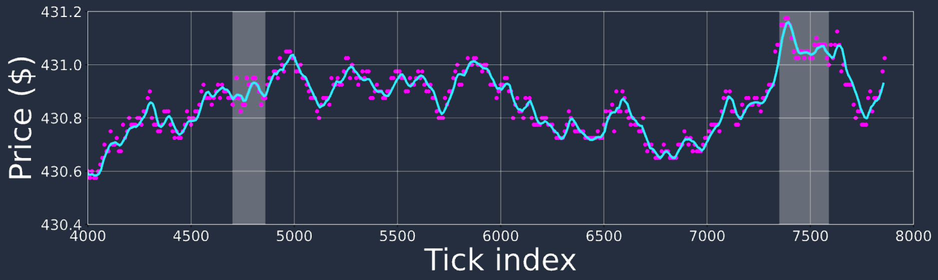
The HFT software stack provides an end-to-end solution for the low-latency finance AI processing on the LightTrader hardware

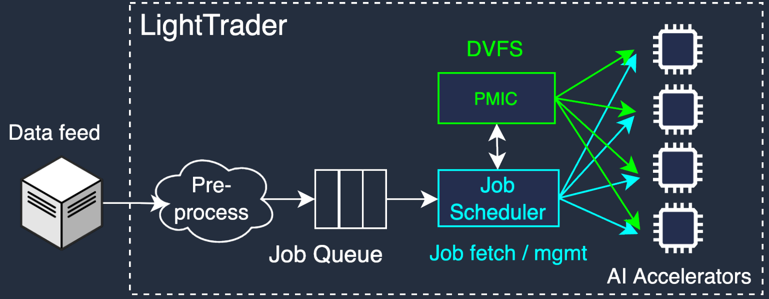
The software stack includes

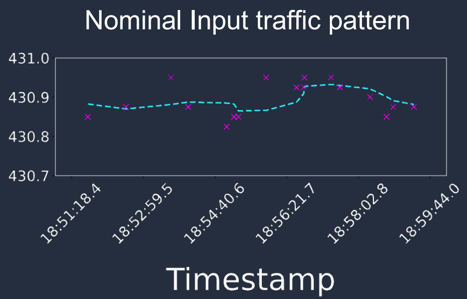
- Compiler
- DL Inference driver
- HFT pipeline driver
- User application interface

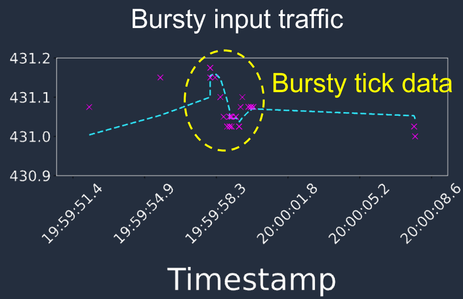
The PCIe Gen 4.0 x16 interface enables a real-time hardware control and monitor

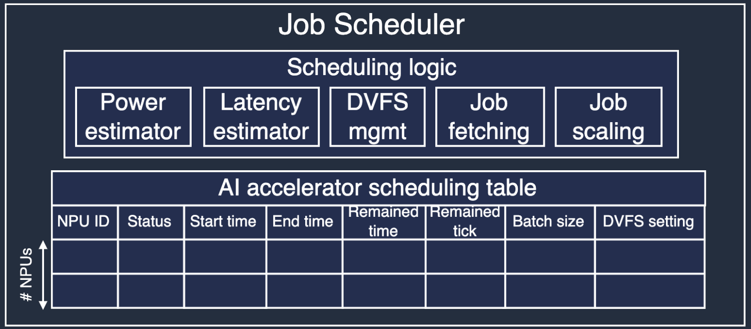
# Job Scheduling enabled by HW features for optimizing throughput






Nominal Input traffic pattern


Bursty input traffic

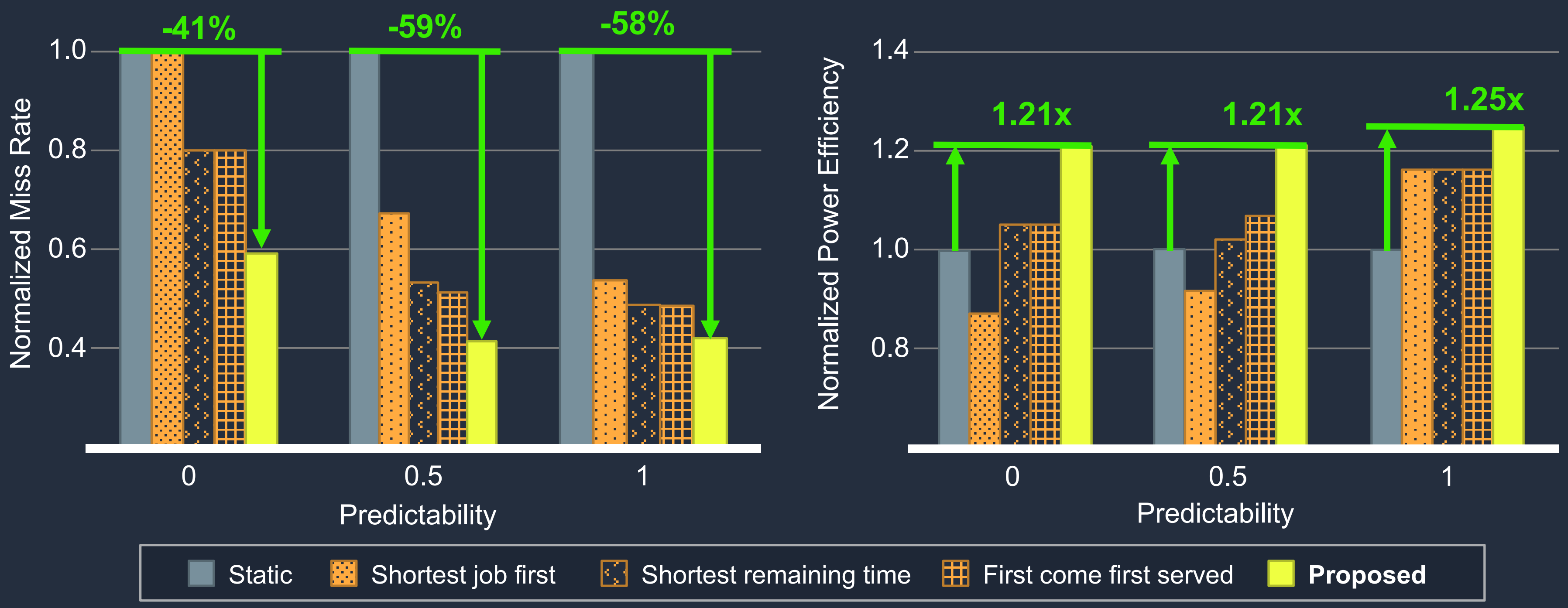


AI accelerator scheduling table

| NPU ID | Status | Start time | End time | Remained time | Remained tick | Batch size | DVFS setting |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| | | | | | | | |

**The advanced scheduling features retain high throughput for tick-by-tick inference jobs even for the bursting input query**

# Miss rate and power efficiency improvements from the intelligent scheduling



The advanced scheduling minimizes the probability of missing input queries

# LightTrader System

## Server-scale Integration

Prototype system integration for AI-based HFT compute node

Integrating eight boards 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 board power

# rebellions_

Revolutionize AI Silicon

hyunsungkim@rebellions.ai

HOT CHIPS

This work was supported by National IT Industry Promotion Agency(NIPA) grant funded by the Korea government(MSIT)
(R-20220319-010567, Development of AI accelerator and its software full stack for edge server system)