DRAVA: AN EVENT-DRIVEN RUNTIME FOR SCIENTIFIC PIPELINES WITH AGENTIC CONFIGURATION OPTIMIZATION

Ahmedur Rahman Shovon, Romain Pereira, Xingfu Wu, Brice Videau, Swann Perarnau
Argonne National Laboratory

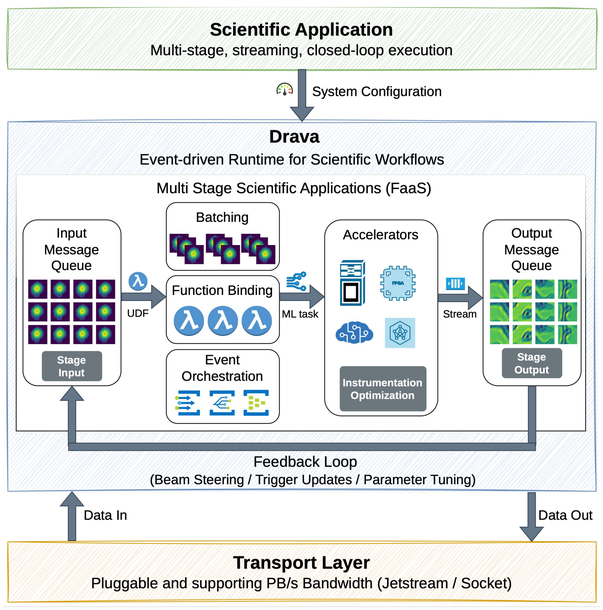
ABSTRACT

Drava is an event-driven runtime for scientific streaming workflows. It uses a lightweight Function-as-a-Service model where application stages are expressed as reusable functions triggered by data events.

Core contributions of Drava are:

- Event-driven **abstraction**
- Transport-aware **orchestration**
- End-to-end **observability**
- Agentic **optimization**

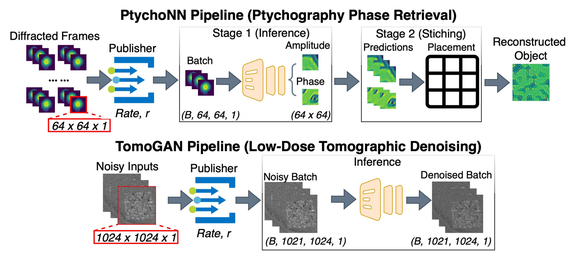
Compared with state-of-the-art PvaPy, Drava handles unpaced input and achieves up to 2.36x higher throughput. Batching improves throughput by up to 2.44x, agentic Bayesian optimization finds a configuration 2.24x faster than best observed manual configuration.



Drava enables event-driven, observable, and optimizable scientific pipelines

MOTIVATION

- APS Upgrade: 500x brighter beams
- PB/s-class detector data rates
- kHz acquisition, real-time feedback
- Manual tuning is brittle



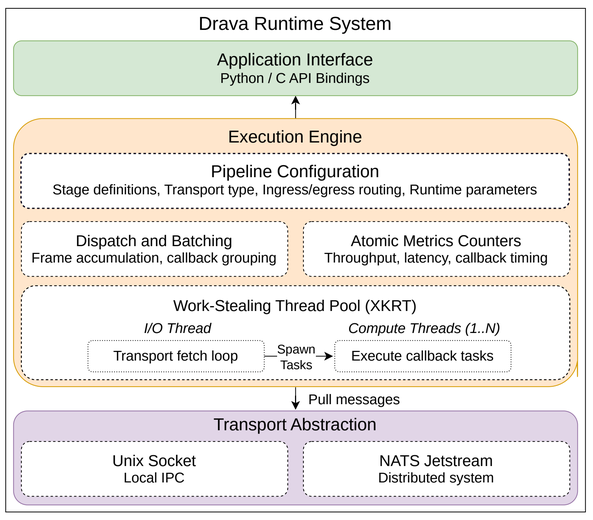
Agentic autotuning of hyperparameter space

Application knobs	Runtime knobs
• Batch size	• Concurrency
• Model params	• Transport
• Stage configs	• Placement

Computation ↔ Data transfer
Combinatorial configuration space

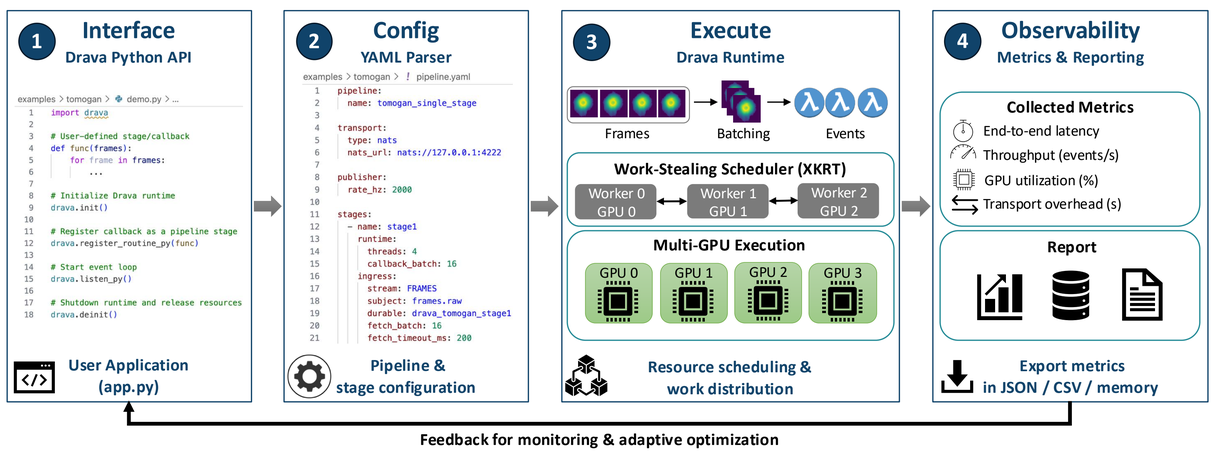
RUNTIME DESIGN

- Layered abstraction model
- Built-in dispatch and batching
- Event-driven programming model
- End-to-end performance metrics

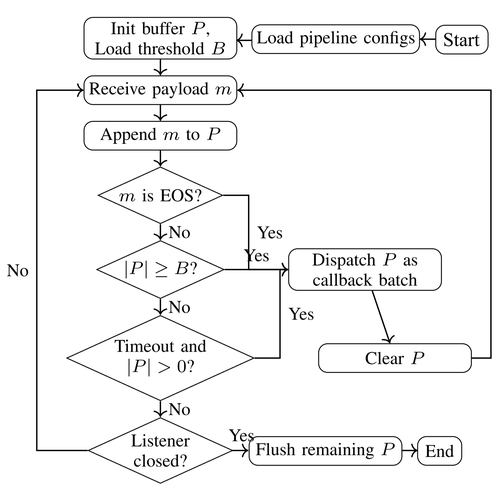


Multi-layered runtime for high-throughput event driven applications

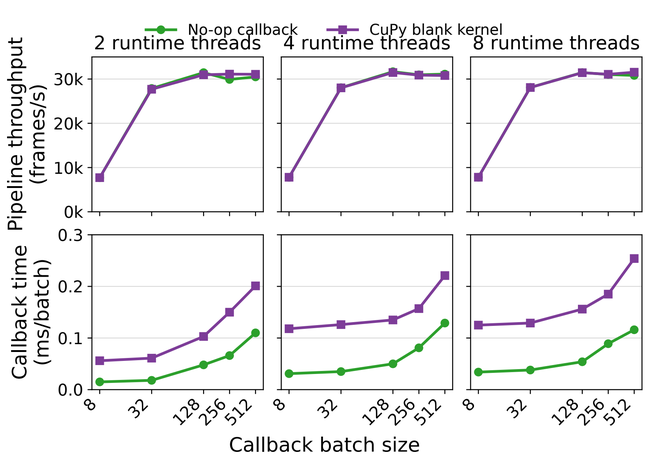
WORKFLOW & BATCHING



Drava application workflow: define, configure, execute, observe



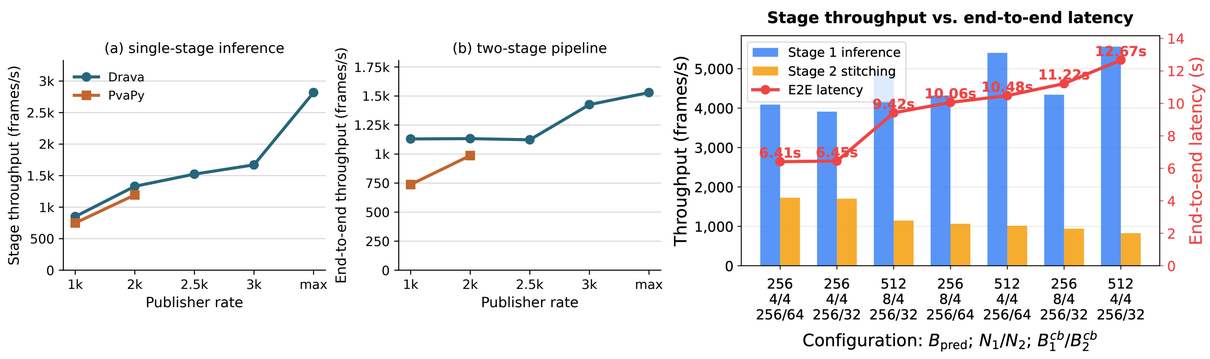
Dispatch: threshold / EoS / timeout



Runtime message-rate ceiling: ~31K frames/s

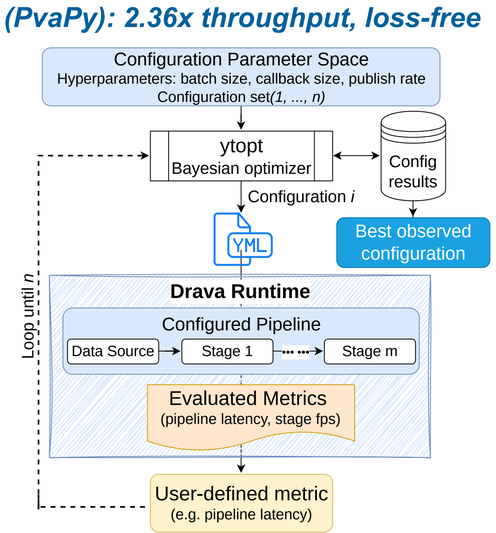
BASELINE & AGENTIC TUNING

- Baseline: PvaPy
- Hyperparameter tuning: Bayesian optimization using ytopt
- Environment: JLSE node from ANL (AMD EPYC 7532, NVIDIA A100 40 GB)
- Applications: PtychoNN (Ptychography), TomoGAN (Tomography)

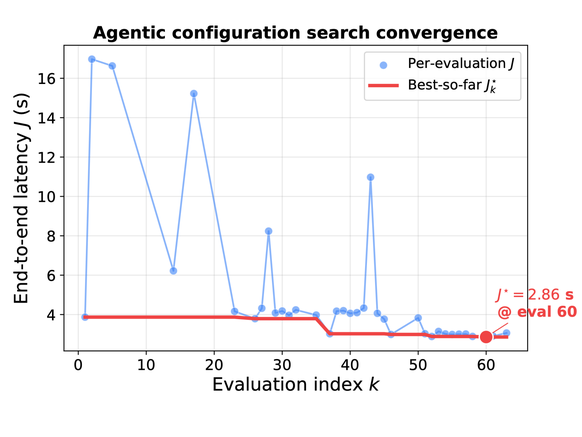


SOTA (PvaPy): 2.36x throughput, loss-free

Throughput vs latency in manual configs

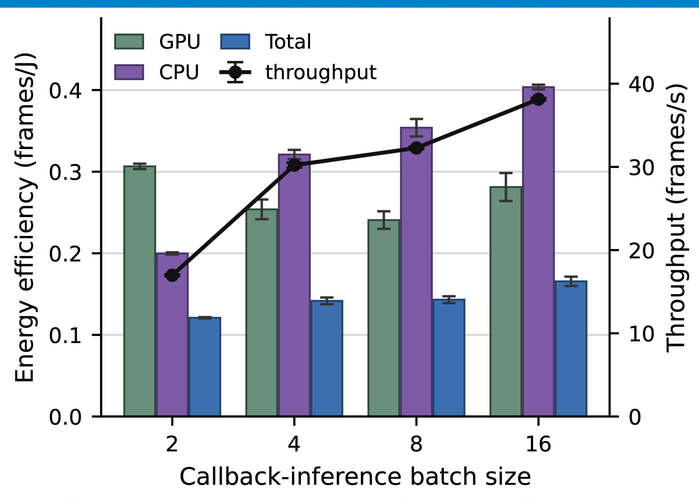


Agent aided configuration search using Bayesian optimizer



Agentic exploration: 2.24x faster, only 0.52% space explored

OBSERVABLE ENERGY EFFICIENCY



Energy efficiency and throughput for TomoGAN inference

CONCLUSIONS & FUTURE WORK

- Event-driven, FaaS-style runtime sustains multi-stage streaming pipelines
- Agentic Bayesian optimization navigates combinatorial config space
- Lock-free observability: per-stage and end-to-end analysis

In future, we will

- Apply dynamic configuration adaptation
- Integrate a high-throughput transport backend
- Explore runtime execution on future accelerators

REFERENCES

- Veseli et al. PvaPy streaming framework for real-time data processing. Synchrotron Radiation, 32(3), 823-836.
- Cherukara et al. AI-enabled high-resolution scanning coherent diffraction imaging. Applied Physics Letters, 117(4).
- Liu et al. TomoGAN: low-dose synchrotron x-ray tomography with generative adversarial networks: discussion. Journal of the Optical Society of America A, 37(3), 422-434.
- PEREIRA et al. XKRT: a Runtime System for Macro-Dataflow Programming on Multi-Devices Architectures. Available at SSRN 6460634.
- Chien et al. Updown: A supercomputer co-designed for scalable graph processing. IEEE Transactions on Parallel and Distributed Systems.

U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

Argonne National Laboratory's contribution is based upon work supported by Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357. This work was supported in part by DOE ASCR Microelectronics project BIA, and in part by DOE ASCR SciDAC RAPIDS. This work was supported by the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

Bia Co-Design for Energy-Efficiency
Argonne NATIONAL LABORATORY