Taurus: An Intelligent Data Plane

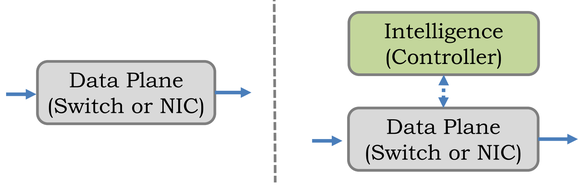
Tushar Swamy, Alexander Rucker, Muhammad Shahbaz,
Neeraja Yadwadkar, Yaqi Zhang, and Kunle Olukotun

Problem Statement

Approaches to managing datacenter networks are either:

Fast yet dumb
(e.g., ECMP)

Slow but intelligent
(e.g., anomaly detection)



Can we get ***fast and intelligent*** approaches by moving the intelligence into the data plane?

Data Plane Expressiveness

- Popular networking DSLs like P4 already comprise of multiple programming abstractions.
- However, none of these are suitable for machine learning tasks.
- We propose adding a new abstraction for machine intelligence: **Map-Reduce**.

Abstraction	Implementation
Parsing	FSMs:
Match-Action	RMT+VLIW:
Scheduling	PIFO:
Map-Reduce (Intelligence)	**Parallel Patterns**

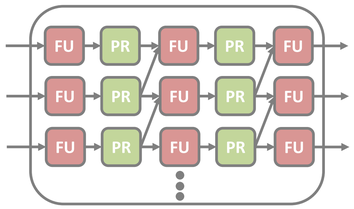
Design of Taurus

We design our hardware around the **Map-Reduce** pattern to support machine intelligence and exploit inherent parallelism for high throughput and low latency execution of learned algorithms.

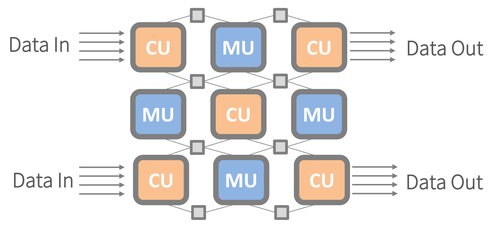
Architecture

A Map-Reduce compute unit consists of:

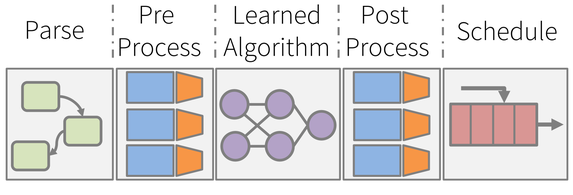
1. Pipelined SIMD lanes of functional units (**FU**) and pipeline registers (**PR**).
2. A reduction network …



Compute Units (**CU**) are interspersed with scratchpad memory units (**MU**) to exploit data locality.



Taurus Pipeline



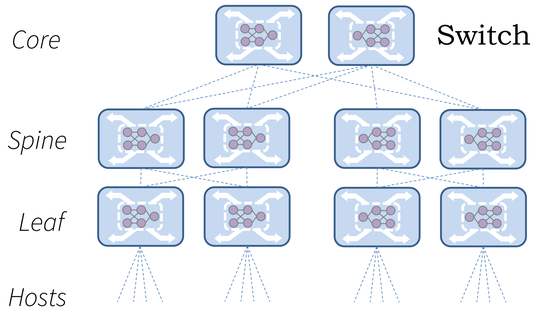
- The Map-Reduce engine is embedded between match-action tables in the ingress pipeline.
- We prototype Taurus using the **Plasticine** architecture and **Spatial** HDL.
- Taurus's peak throughput ranges from 512 Gbps – 12 Tbps depending on packet and model sizes.

Taurus in the Datacenter

Datacenter operators can run Taurus both inside the network switches as well as at the end-host NICs.

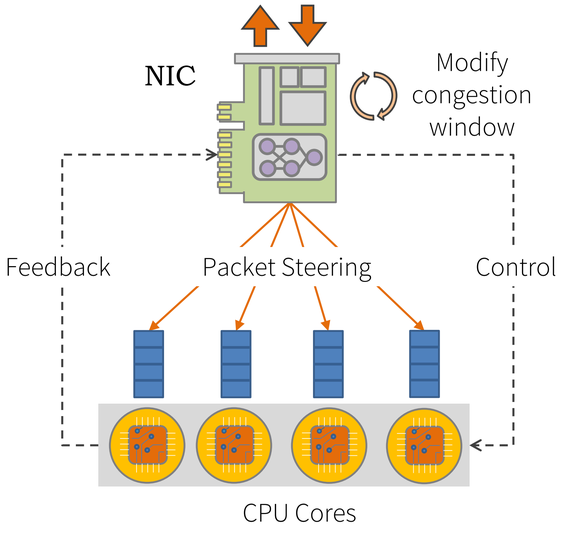
Taurus in Switches

- High throughput allows switches to apply learned functions to **every packet** in the network.
- Enables intelligent applications like **anomaly detection, routing,** and **load balancing** using learned functions that operate at line rate.



Taurus in Hosts' NICs

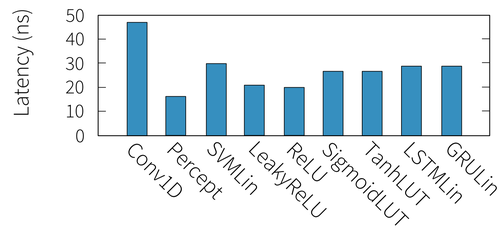
- Taurus can implement end-host functions for applications like **congestion control (Indigo)** and **RSS (Shenango).**
- Learned algorithms with feedback allows for smarter core and packet scheduling.



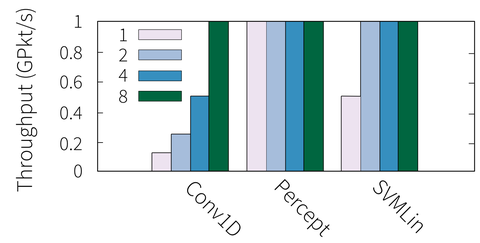
Evaluation

Microbenchmarks

- Latency across different microbenchmarks:



- Unrolling factors required to hit 1 GPkt/s:



App. Benchmarks

- Overheads are calculated relative to a 300 mm² chip with 4 pipelines each drawing an estimated 25 W.
- Anomaly Detection:
  The models are fully unrolled to meet typical switch line rates (512 Gbps – 12 Tbps).

		Area		Power	
Model	Lat (ns)	mm²	+%	mW	+%
SVM	68	4.59	6.1	263	1.1
DNN	362	8.80	11.7	506	2.0

- Indigo Congestion Control
  The model is unrolled to meet typical NIC line rates (40 Gbps – 96 Gbps).

		Area		Power	
Model	Lat (ns)	mm²	+%	mW	+%
LSTM	380	17.73	23.6	1018	4.1