OSF OpenInfra Labs
Project Caerus Kickoff Meeting

August 26, 2020

Agenda

- Opening Remarks (Jonathan Bryce, Hui Lei)
- Vision, Scope, and Opportunities (Hui Lei, Theodoros Gkountouvas)
- Related Work at BU and NEU (Orran Krieger, Peter Desnoyers, et al.)
- Open Discussion on Research and Development Plan (all)

OSF, OpenInfra Labs, and Project Caerus

- The OpenStack Foundation (OSF) has expanded its scope to address open infrastructure in general
- In addition to OpenStack, OSF currently hosts OpenInfra Labs and several other active projects
- Project Caerus is a sub-project in OpenInfra Labs

OpenStack Foundation (OSF) Open Infrastructure Projects

- Airship: automated cloud lifecycle management
- Kata containers: secure, lightweight virtualized containers
- **OpenInfra Labs: cross-stack integration and optimization**
- OpenStack: software-defined infrastructure
- Starling X: edge computing infrastructure
- Zuul: CI/CD platform across multiple systems/repos

OpenInfra Labs Sub-projects

- OperateFirst: open source cloud operations
- **Caerus: compute-storage coordination**
- ESI: secure and elastic bare-metal infrastructure
- Wenju: integrated development of AI applications

Open Research Collaboration

- Boston University
  - Prof. Orran Krieger
- Columbia University
  - Prof. Ken Ross
- Northeastern University
  - Prof. Peter Desnoyers
- Ohio State University
  - Prof. Xiaodong Zhang

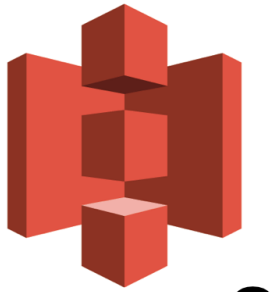
Project Caerus:
Optimizing the Big Data Ecosystem

Hui Lei

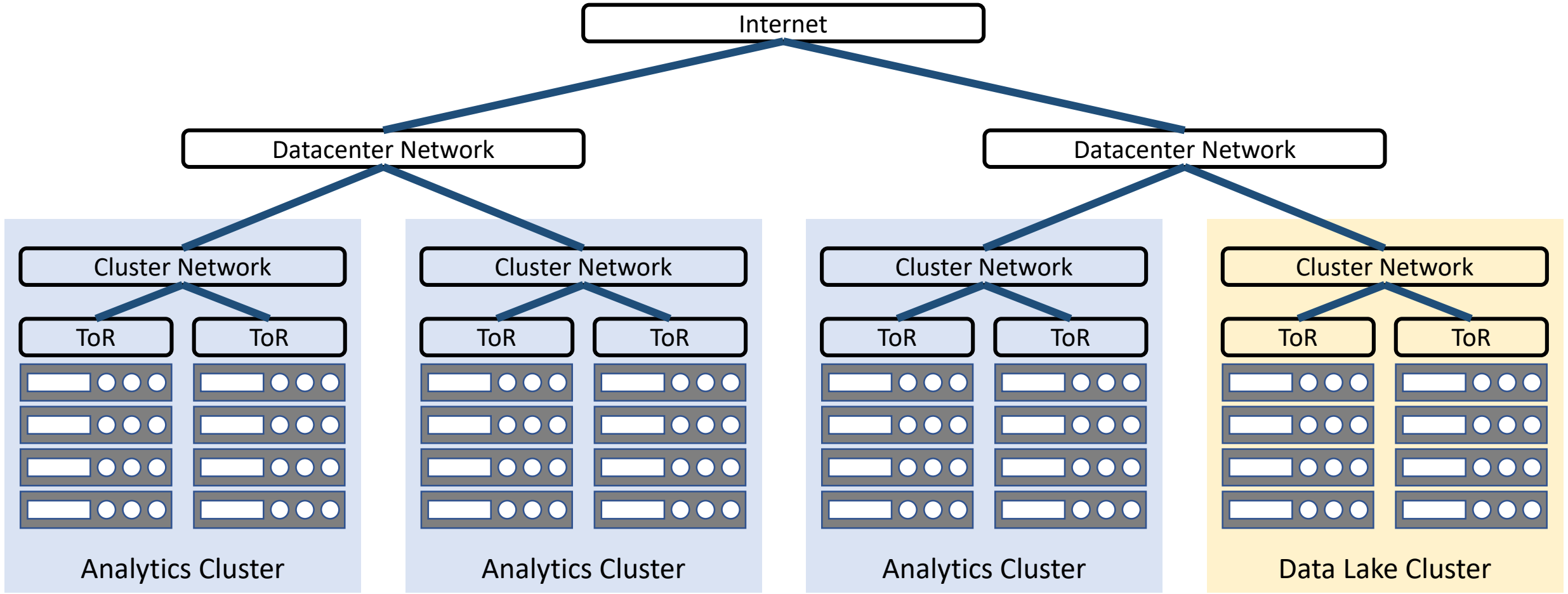
The Big Data Ecosystem

Analytics Engines:

Data Lakes:



Common Pattern: Disaggregated Compute and Storage



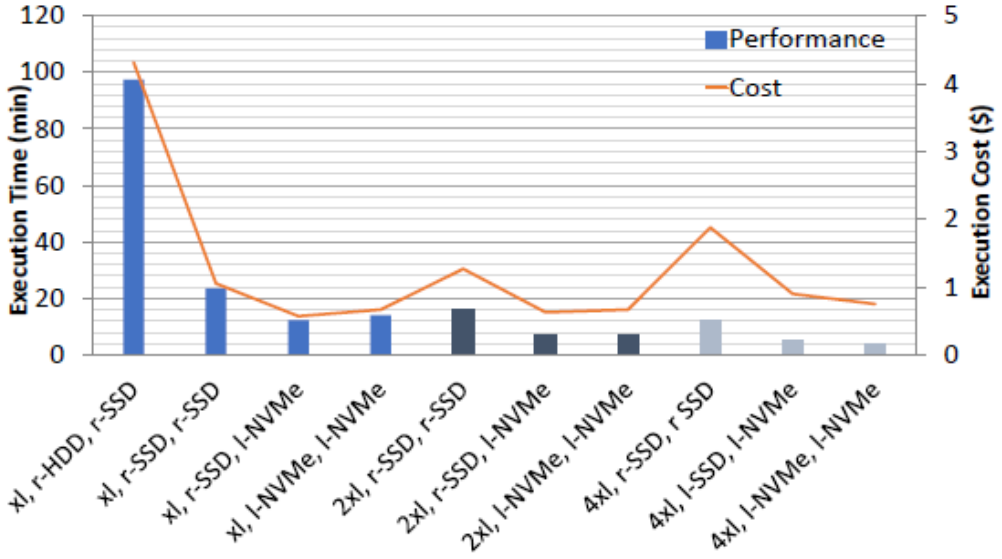
Benefits

- Data sharing via centralized collection and management
- Elastic scaling of the compute and storage infrastructure
- Improved utilization of datacenter resources
- Easy adoption of compute and storage innovations

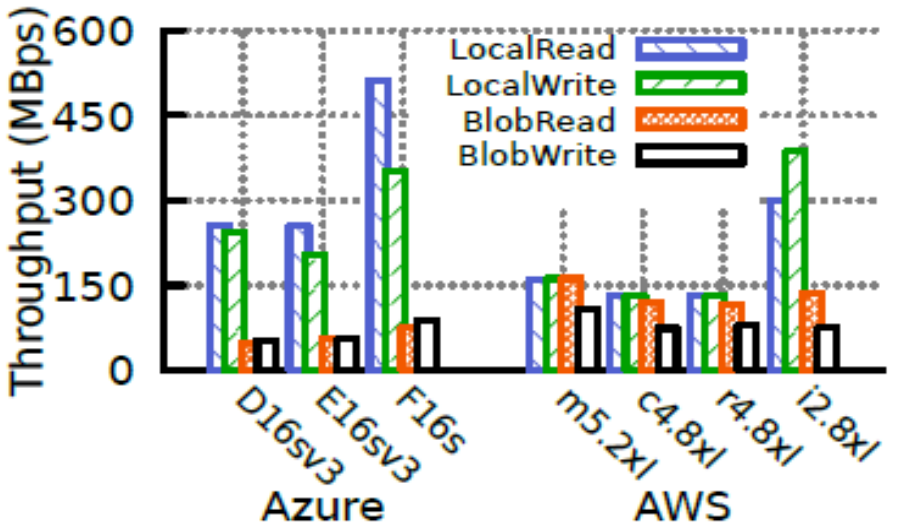
Challenges:

- Data access latency and performance variation due to physical distance
- Constrained data throughput due to multiple levels of oversubscribed networks
- I/O bottleneck due to the limited speed of disks used for very-high-capacity storage
- Impeded performance of data-parallel analytics due to the memory wall

Performance Disparity Between Local and Disaggregated Storage

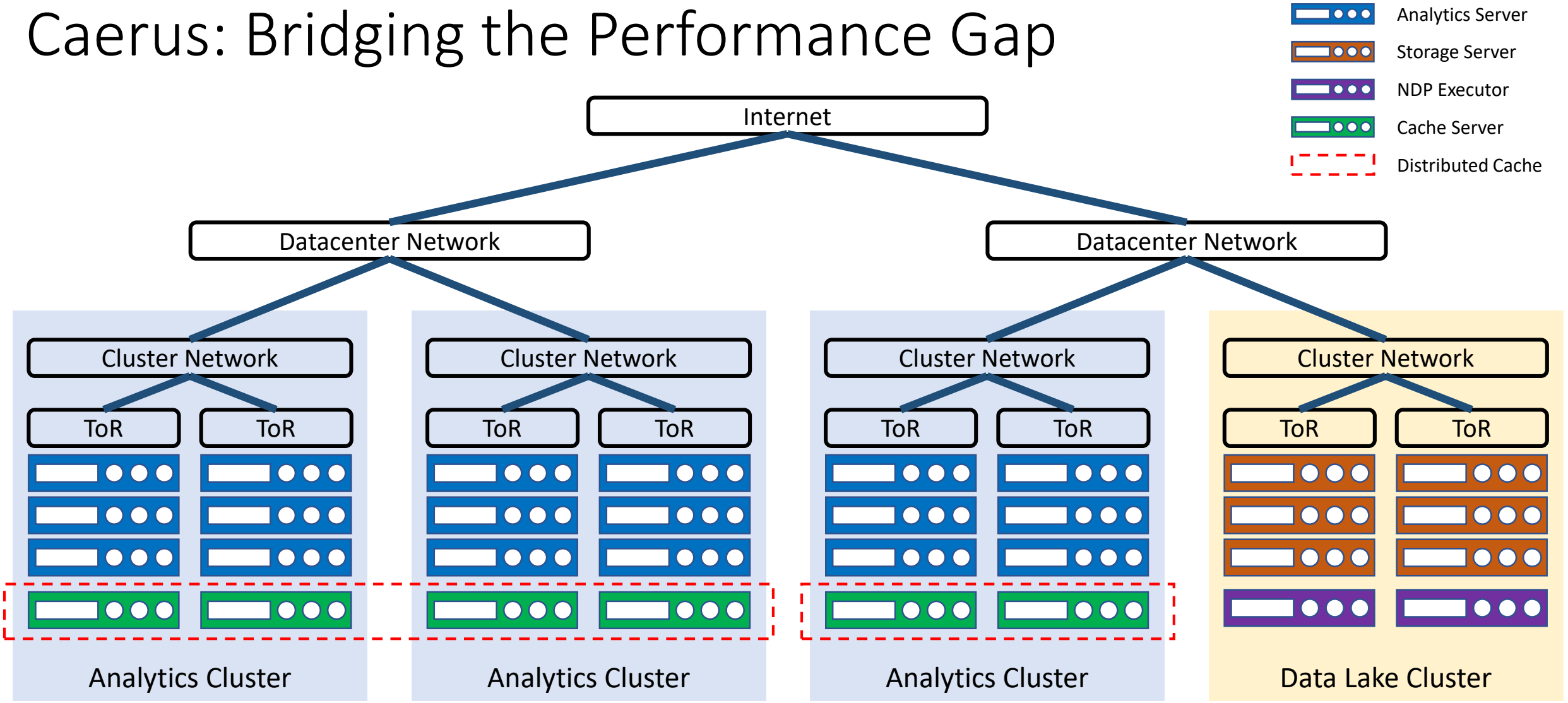


Stanford Selecta (ATC ’18): Comparison of execution time and cost for TPC-DS query 64 on various EC2 VM and storage configurations



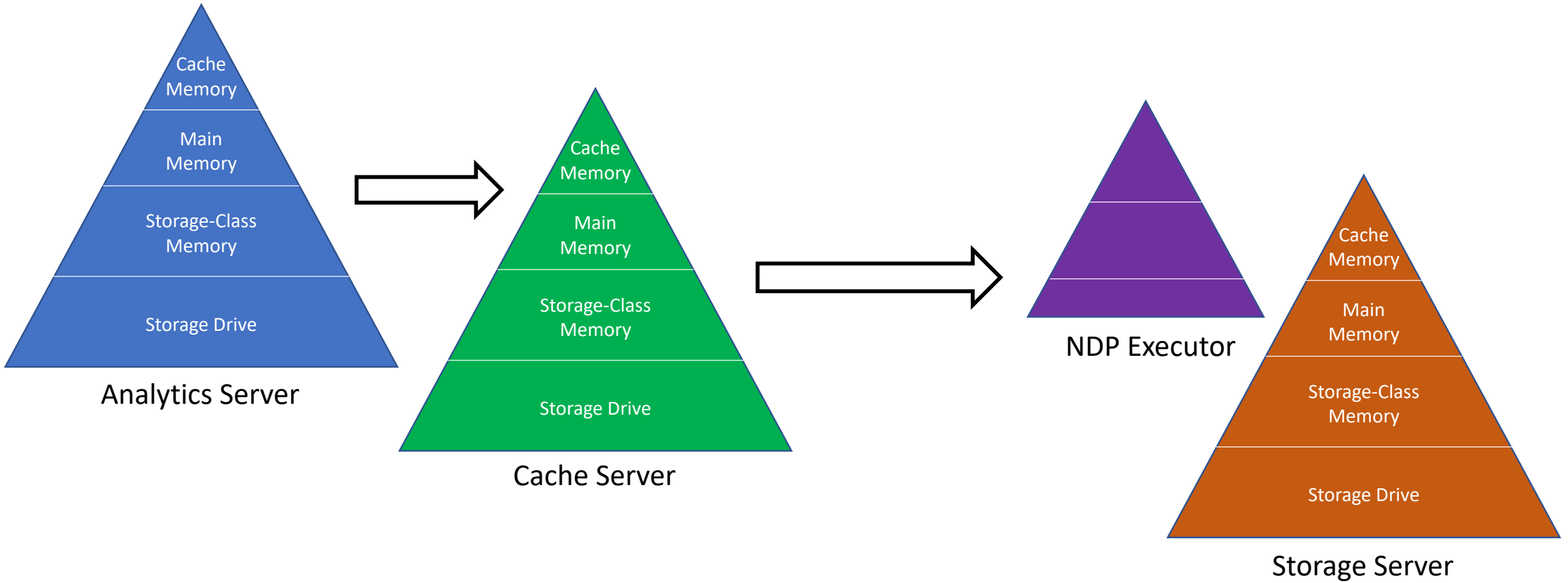
Microsoft Netc (SOCC ‘18): Comparison of I/O throughput for local and remote stores on Azure and AWD for different VM types

Caerus: Bridging the Performance Gap



- **Near-data processing:** opportunely pushing a broad array of computation close to data
- **Semantic caching:** distributed, multi-modal, and workload-aware caching substrate
- **Holistic optimization:** coordination of all optimization techniques and across analytics engines
- Targeting full utilization of CPUs/GPUs for data-parallel workloads and 3x-5x reduction in application execution time

The Memory Hierarchy



- At a macro level, the memory hierarchy extends from the analytics server and cache server to the NDP executor and storage server
- At a micro level, each server has memory that spans cache memory, main memory, storage-class memory and hard disk
- In principle, NDP and caching can be carried out at any level of the memory hierarchy
- Caerus focuses on the layers of the memory hierarchy that matter the most for massively data-parallel analytics
  - NDP: Primarily the NDP Executor's CPU and, to a lesser extent, the Cache Server's CPU
  - Semantic Caching: Primarily the Cache Server's storage drive and, to a lesser extent, the Storage Server's storage drive

The Scope of Near-Data Processing

- **What to pushdown**
  - Simple query operations: select, project, aggregation
  - Complex query operations: join, group-by, top-K, search
  - AI functions: K-mean, classification, shuffle
  - User-defined functions: feature extraction, format transformation, metadata enrichment, thumbnail generation, checksum calculation, data de-identification
- **How to pushdown**
  - Declarative: SQL-like query language, DAG-based specification
  - Imperative: container image, serverless function
- **How to execute**
  - NDP executor integrated into a storage server (e.g., object storage, database)
  - NDP executor as an external server close to a storage server (e.g., HDFS)
  - Hardware and software acceleration: SIMD processor, GPU accelerator, zero-copy CSV file parser
- **Where to pushdown to**
  - Pushing processing from the analytics server to the NDP executor and the cache server

NDP Industry Landscape

Software System	Analytics Engine	Storage System	What to Pushdown	How to Pushdown	How to Execute
MinIO	Any, with custom optimization for Spark (Spark-Select)	MinIO object store with S3 API	Simple query ops, UDFs	S3 Select API	Object store server with acceleration for limited ops
Amazon S3 and Glacier	Any	S3 object store	Simple query ops	S3 Select API	Object store server
Ceph	Any	Block/File/Object abstraction on Ceph object store	UDFs	Extended API for UDF deployment on the storage server	LUA VMs on storage side talk to object storage daemons (OSDs) via extension interfaces
OpenStack Swift	Any	Swift object store	UDFs, simple query ops	Storelets extension to Swift	Storlet containers on storage side support UDFs and simple SQL queries
Spark	Spark	Compatible data stores	Simple query ops	Spark filter() and select() methods	Storage-specific
Caerus	Multiple	Object stores, HDFS	Simple and complex query ops, AI functions, UDFs	Queries, DAGs, container images, serverless functions	Standardized NDP executor internal or external to storage server, with hw and sw acceleration

NDP Research Innovations

University	Project Name	Conference	Analytics Engine	Storage System	What to Pushdown	How to Pushdown	How to Execute
University of California, Santa Cruz	SkyhookDM	FAST 2019	PostgreSQL	Ceph	Simple query ops, UDFs	Custom client interface	Ceph UDF mechanism
Reutlingen University & TU Darmstadt Germany	nativeNDP	ADBIS 2019	R Platform	Ceph	R-native operations	Ceph UDF mechanism	Ceph UDF mechanism
University of Wisconsin-Madison & MIT	PushdownDB	ICDE2020	Homegrown PushdownDB	S3	Both simple and complex query ops	S3 Select interface	S3 Select mechanism with reimplementation of the more complex DBMS operations
University of California, Irvine	Catalina	PDP 2019	Any analytics application	Catalina (custom SSD with multicore processor running Linux OS)	Special example: image similarity search using MPI	MPI slaves embedded in storage devices	In-storage application processor along with FPGA accelerators

- Modification of the analytics engine to take advantage of NDP
- Support for the pushdown of more complex operations
- Acceleration of NDP on storage side

The Scope of Semantic Caching

- **Distributed caching:** a pool of cooperative caching nodes serving the analytics clusters
  - Decentralized vs centralized management
- **Multi-modal caching:** caching data of different modalities and stages
  - Source data (objects, files, blocks), intermediate results, partition metadata, re-partitioned data
- **Predicative caching:** prefetching and pre-computing data ahead of demand
  - Strategies on when and what to pre-compute, prefetch and evict
- **Workload-aware caching:** caching decisions driven by high-level semantics
  - History-based vs hints-based approaches
- **Multi-tiered caching:** cache tiering resulted from the multi-node memory hierarchy
  - The persistent memory on the caching and storage servers are most critical for data-intensive analytics

Semantic Caching Industry Landscape

Software / Feature	Storage System	Distributed Cache Management	What Is Cached	Predictive Caching	Workload Awareness	Cache Location
Spark Cache	Multiple	Autonomous	Source data, intermediate data	No	DAG-aware	Analytics side
Azure SparkCruise and Hyperspace	Multiple	Centralized	Source Data, Intermediate Data, indices	No	Yes	Analytics side
Data Skipping in IBM Cloud SQL Query	IBM Cloud Object Store	Centralized	Indices	Prefetching	No	Storage side
D3N Cache in Ceph RADOS Gateway	Ceph	Decentralized	Source data	No	No	Analytics side
Alluxio	Multiple	Centralized	Source data	No	No	Analytics side
Caerus	Multiple (incl. Object Stores and HDFS)	Centralized	Source data, intermediate data, metadata, re-partitioned data	Prefetching, pre-computing	DAG-aware	Analytics side, storage side

Semantic Caching Research Innovations

Organizations	Project Name	Conference	Storage System	Distributed Cache Management	What Is Cached	Predictive Caching	Workload Awareness	Cache Location
Texas, Microsoft Research	INSTalytics	FAST 2019	HDFS	Centralized	Source Data, Static replicas with different partitioning	No	Selection Operation	Storage side
MIT CSAIL, Microsoft	Amoeba	SoCC 2017	HDFS	Centralized	Source Data, Dynamic replicas with adaptive re-partitioning	No	Selection Operation	Storage side
UC Berkeley	Partitioning for Aggressive Data Skipping	SIGMOD 2014	HDFS	Centralized	Features with supporting partitioning	No	Selection Operation	Storage side
Hong Kong University	LRC, LERC	INFOCOM 2017	Multiple	Centralized	Intermediate Data	No	Spark DAG	Compute side
BU, NEU, State Street	Kariz	HotStorage 2019	Ceph	Centralized	Source Data	Prefetching	Spark / Pig / Hadoop DAG	Compute side

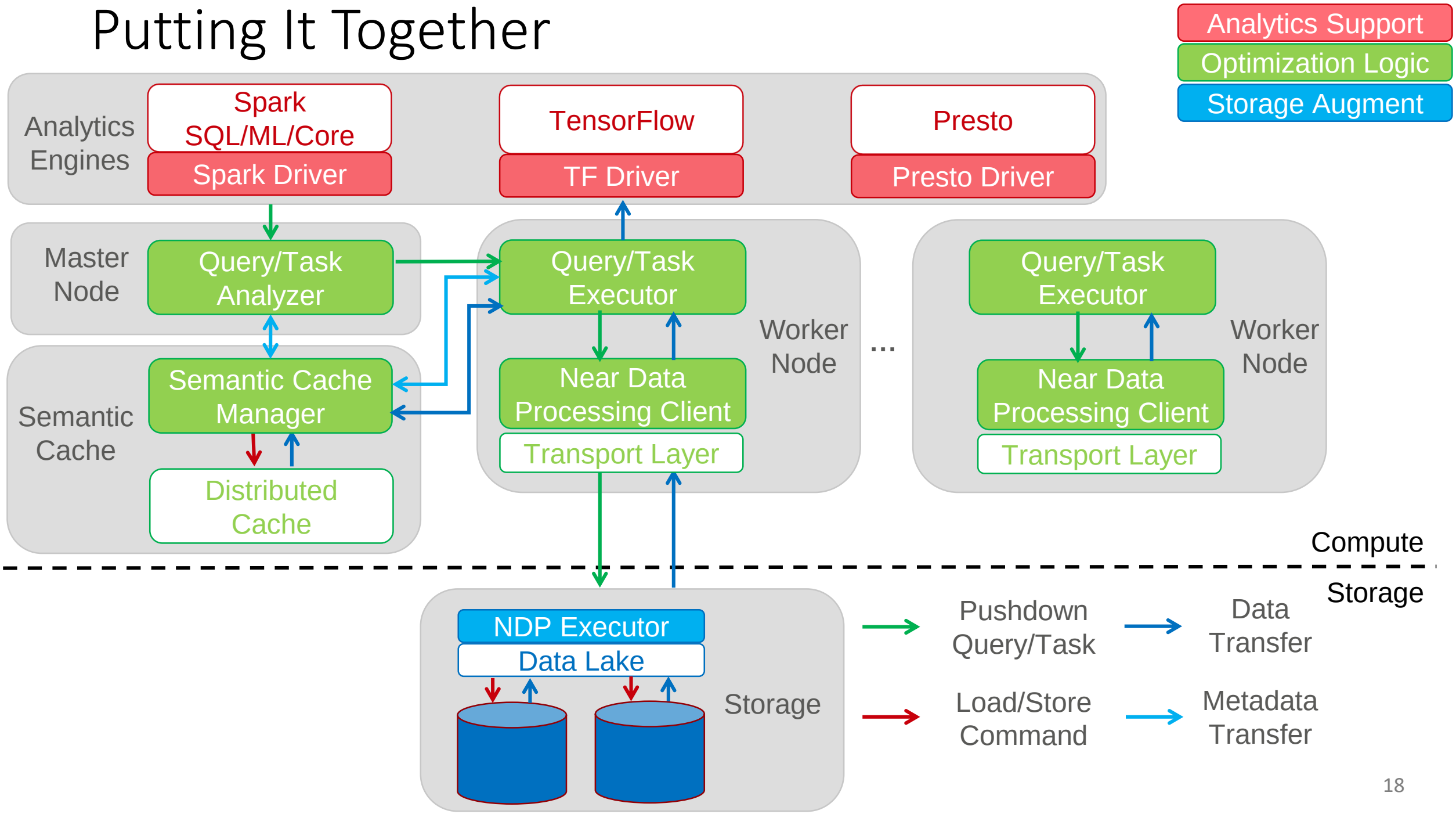
- Storage-side replication with differing partition schemes
- Advanced data skipping techniques
- Workload-aware caching and prefetching

The Scope of Holistic Optimization

Holistic optimization is required across

- Multiple semantic-caching techniques
- Distributed caching nodes
- Caching and NDP
- Multiple workloads
- Multiple analytics engines

Putting It Together



Caerus Differentiations

- Pushdown of a broad range of operations: simple and complex query operations, AI functions, UDFs
- Caching of a variety of content: source data, intermediate results, metadata, and re-partitioned data
- Prefetching and pre-computing data in addition to caching recently used data
- Leveraging infrastructure resources on both analytics-side and storage-side
- Using workload semantics to drive all NDP and caching decisions
- Exploiting hardware and software acceleration for decisions, NDP, and pre-computing
- Holistic optimization and automated configuration across many NDP and caching techniques
- Extensible architecture to accommodate different analytics engines and storage systems