

Developing a Data-Centric Ecosystem for the Big Data Revolution

Edmund Yeh and Lixia Zhang

Problems

- Observation: many big data application domains exhibit similar set of problems
 - LHC high energy physics
 - Climate
 - LIGO, medicine, ..
- System challenges: data storage, indexing, distribution, security, privacy
- Today: domain experts are dealing with these systems problems
 - Incremental solutions
 - Developed in isolation, replicated efforts

Root Cause

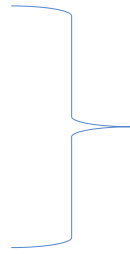
- There exist a gap between what applications need and what the existing systems offer
- Current computer systems/networks focus on addresses, processes, servers, connections
 - Consequently existing security solutions focus on securing data containers and delivery pipes
- Applications care about data

Solution Directions

- Data-centric approach to system and network design
- Providing system support through the whole data lifecycle:
 - Data production: naming, securing data directly
 - Delivering data using names enables scalable data retrieval
 - multicast delivery
 - in-network caching
 - automated joint caching and forwarding
- Common framework to support different application domains
- Cross-cutting theme

X-Centric Designs

- Computing-centric
- Service-centric
- User-centric



Driving design focus

- toward semantic meaning
- away from nodes/machines

Communication semantics: embedded in data

Networking: delivering bags of bits → data-centricity encompasses all

Convergence Toward Data-Centricity

- Emerging Pub-sub paradigm in application development
 - e.g. MQTT
- Data-centric approach already appear at different system levels
 - LABIO: New approach using label to represent data in I/O systems
 - Data warehouses

Research Challenges

- Data-centric systems/networks: naming data
 - Instead of naming locations
- Namespace design offers potential opportunities
 - Hierarchically structured semantic naming
 - Provides context, specifies data operations, enables policies
- Namespace design raises great challenges
- Challenges in integrated system/network, memory/storage/communication design