Streaming Data Analysis on the Wire (AoW)

NYSDS 2016

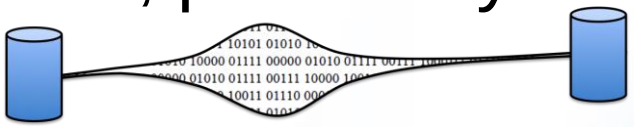
Dimitrios Katramatos (BNL)
Jin Xu (SBU)
Jiayao Zhang (SBU)
Shinjae Yoo (BNL)
Meng Yue (BNL)
Kerstin Kleese van Dam (BNL)

Outline

- Motivation
- Concepts
- Potential use cases
- Current solutions
- Challenges
- Research directions
- Work in progress and preliminary results
- Conclusions and future work

Motivation

- In the Big Data era, lots and lots of data can be found at any moment in transit, potentially more than what is in storage



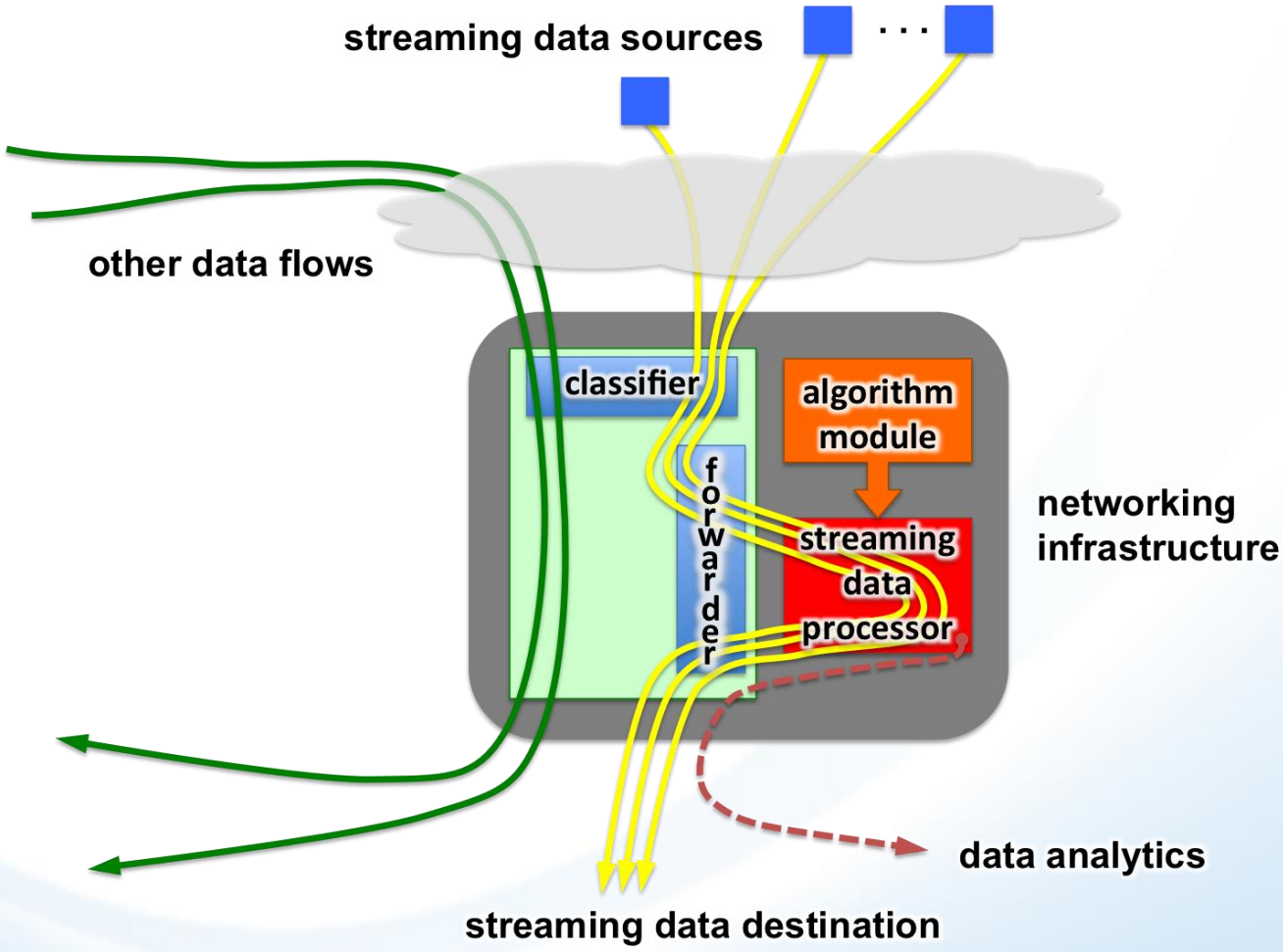
- Is it feasible to devise a framework for data analytics on the wire, i.e., utilizing capabilities of the network infrastructure?
- Early processing provides real-time/near real-time information that can be used to speed up the decision processes

Concepts

- Network infrastructure includes mechanisms that can be programmed to recognize specific data flows based on given criteria
- Flows are then intercepted and transparently forwarded to processing subsystem(s) where data is subjected to desired processing and information is extracted
- Processed data (original or transformed) is ultimately forwarded back to its original destination

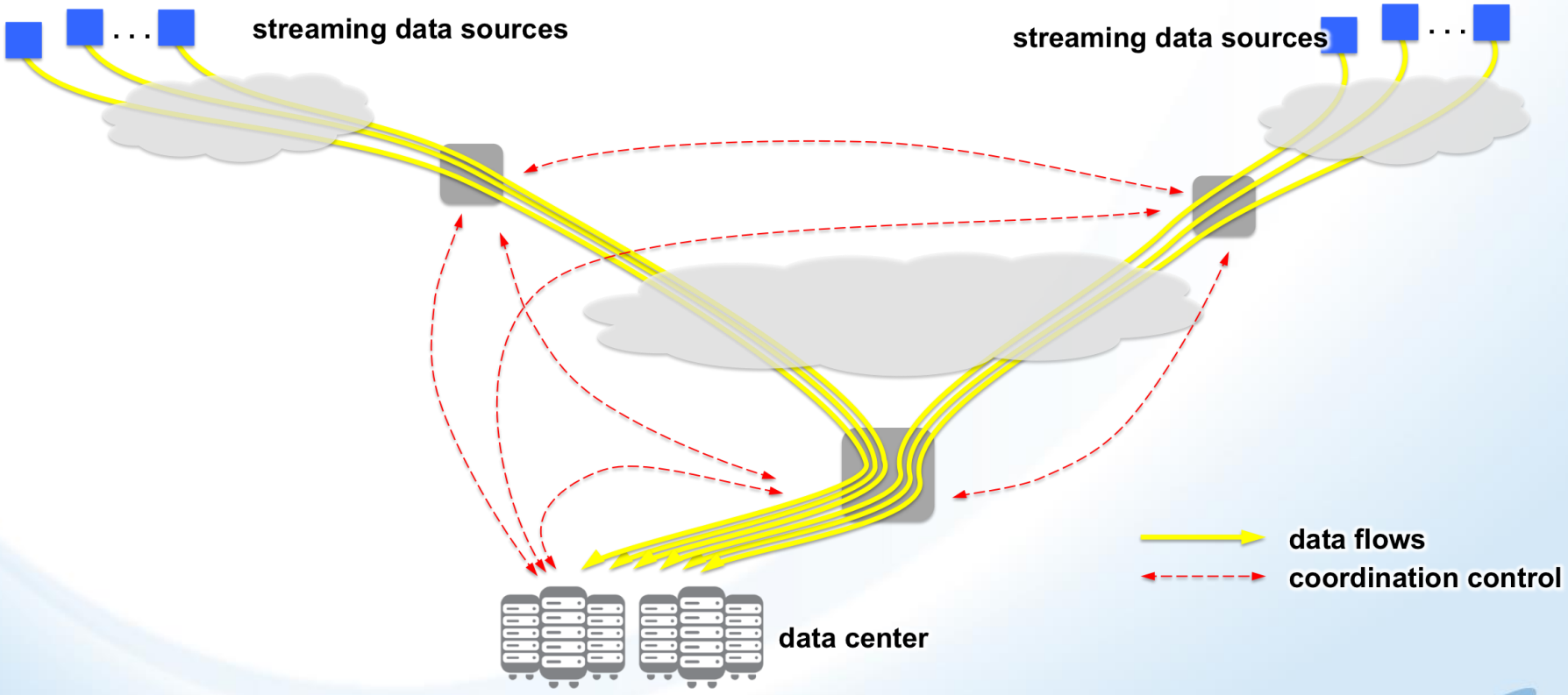
Concepts cont.

Using a strategically located device…



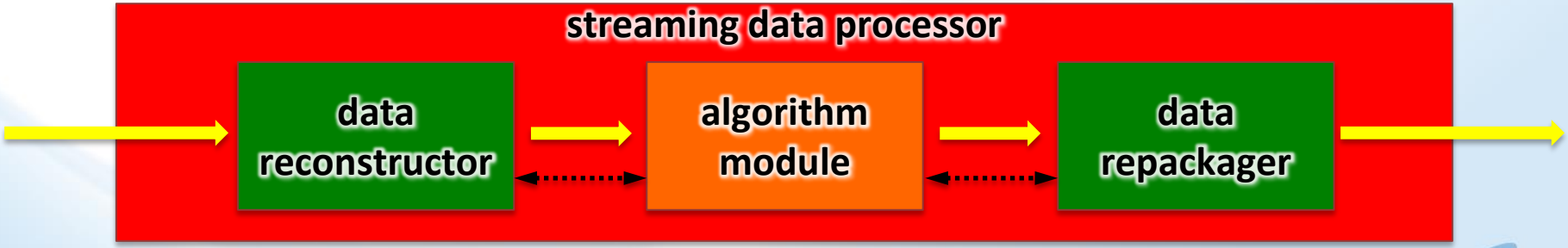
Concepts cont. (2)

… but also consider using multiple coordinated devices at multiple strategic locations, creating a form of distributed computer with many potential applications



Concepts cont. (3)

- Reconstruct portion of data set from streaming packets as algorithm needs
- Algorithm module communicates needs to a reconstructing module through an API
  - Extraction of data analytics does not necessarily affect original packet stream…
  - … but data transformation may require transformed data be “repackaged” into packets in the way the recipient expects



Potential Use Cases

- Scientific
  - Pre-processing during/after data acquisition
    - Early decision making
  - Simple streaming data transformations
  - Statistical information extraction
  - Sensor network data analysis (e.g., distributed solar irradiance prediction, Smart Grid Phasor Measurement Unit and Smart Meter data, DARPA SIGMA security sensor networks)
  - Internet of Things (IoT)

Potential Use Cases cont.

- Commercial
  - Real time processing before arrival at company's data center for decision making
  - Network provider offers value-added processing services to subscribers lacking such capabilities
  - Cybersecurity: go beyond typical firewalls, IDS, etc. in detecting attack patterns and responding
  - IoT

Current Solutions

- Mostly business oriented:
  - Cybersecurity
    - Firewalls
    - Deep Packet Inspection / Processing (DPI/DPP) Systems
    - Intrusion Detection Systems
  - Customer data analytics
- Special hardware and software
  - Vendor-specific
- Software Defined Approaches
  - Also mainly business-oriented
  - Interesting concepts
    - Network Function Virtualization (NFV)
    - Service Function Chaining (SFC)
  - SDN controller software maturity level?

Challenges

- Special-purpose vs. general-purpose
  - User-defined processing
  - Hardware/software is expected to need modifications
- Performance – how much penalty?
  - Additional processing adds overhead
  - Hardware limitations affect both special hardware and SDN solutions
- Algorithms – what can be run?
  - Streaming algorithms with low overhead
  - Loading/distribution – static/dynamic programmability
  - Distributed problems
    - Multi-site/multi-device coordination
    - Scalability

Research Directions

Two directions with equal weight

- Networking
  - Vendor hardware
    - Deep Packet Inspection (DPI) and Deep Packet Processing (DPP)
    - Big IP F5 systems
    - Tap traffic, process out of band in external system
  - Can we influence future vendor designs?

Research Directions cont.

- Prime case for SDN
  - Network Function Virtualization (NFV) and Service Function Chaining (SFC)
    - Need to modify/augment SDN controllers?
  - Augment virtual switch capabilities?
  - White box switches (e.g., Pica8) with customized OS?

Research Directions cont. (2)

- Algorithms
  - Extract streaming data analytics, and/or
  - Transform streaming data
  - Low overhead to match capabilities
  - Examples
    - Outlier detection
    - Approximated summary statistics
    - Lightweight dimensionality reduction using problem characteristics
    - Batch supervised/unsupervised learning
    - Adaptive supervised/unsupervised learning

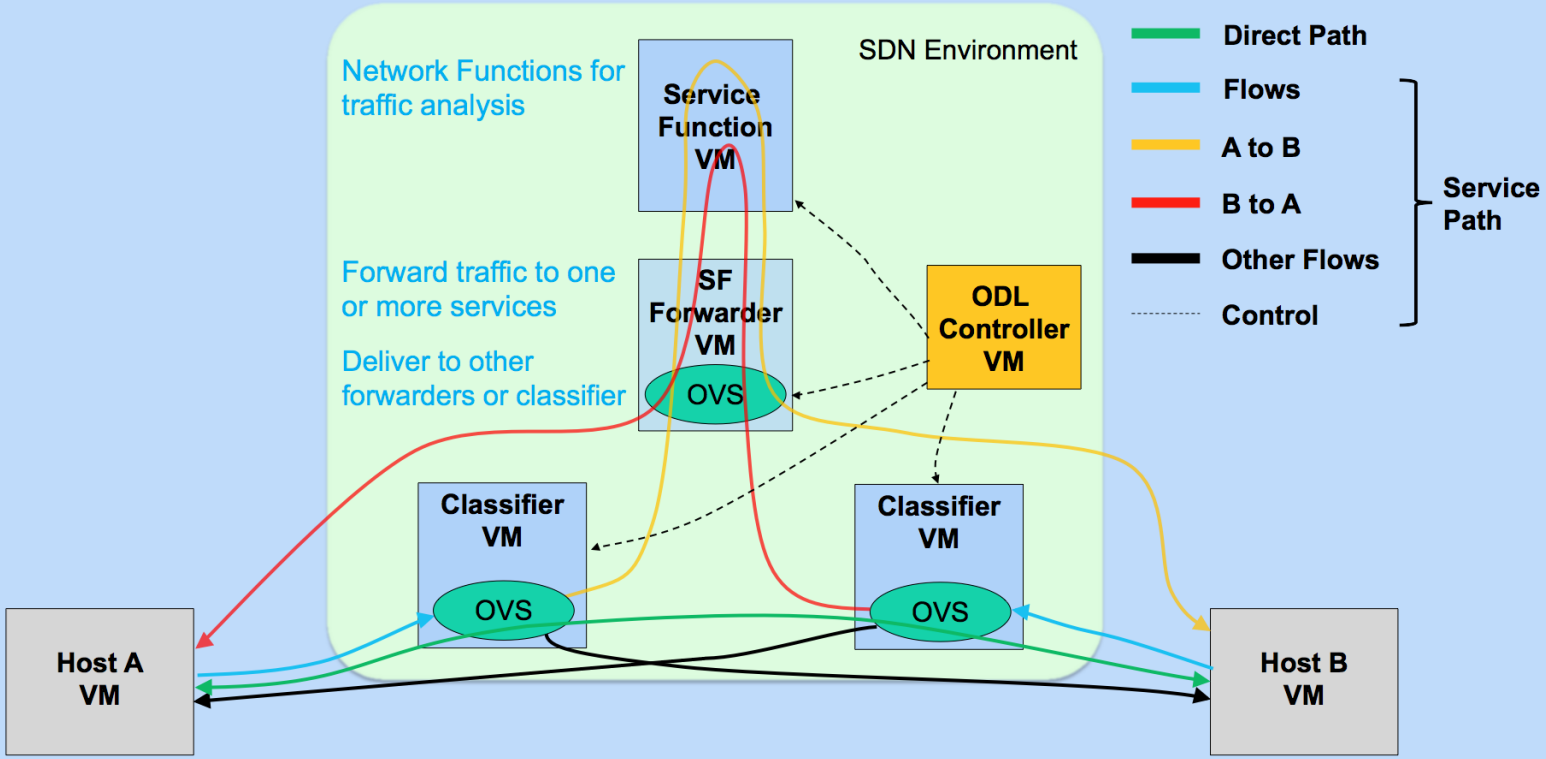
Work in Progress

- Networking
  - Looking into what can be achieved through SDN, especially through SFC
- Algorithms
  - First use case: Smart Grid load forecasting

Exploring the SDN Aspect

- Emulation environment based on SFC tutorial using the OpenDaylight SDN controller and Open vSwitch (OvS) instances
- All components run in VMs, SFC components are configured and/or implemented with python scripts (the service function itself is a python script)
- Service function is “hello world”-style
  - But does do a couple of things, i.e., updates the Network Service Header (NSH), logs packet header

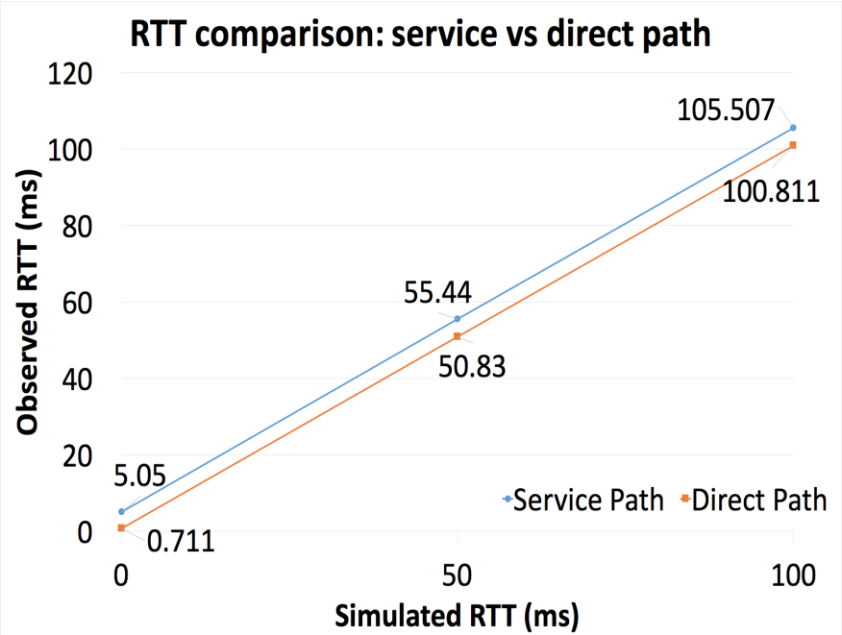
Exploring the SDN Aspect cont.

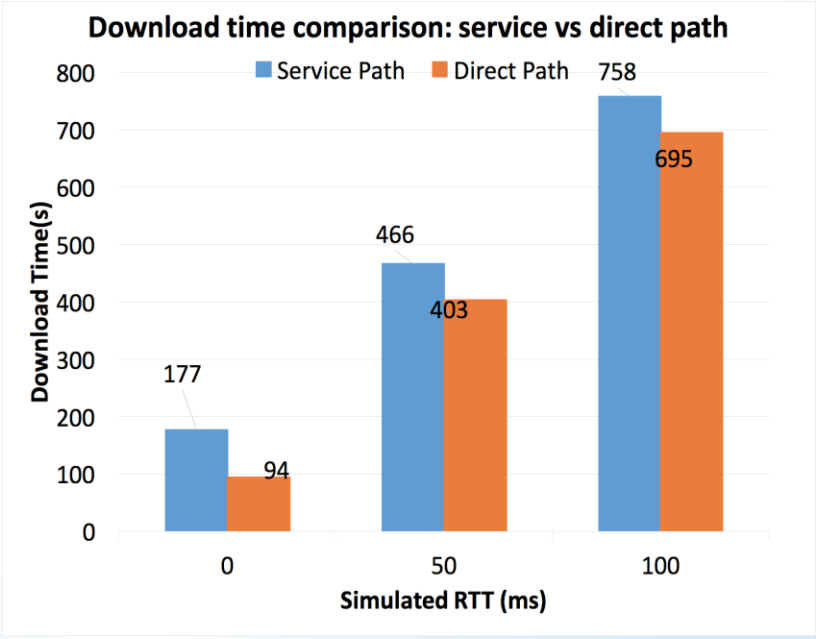


Experimental Setup
Latency of Service vs. Direct Path

Exploring the SDN Aspect cont. (2)

- Tested mechanism with pings, simple file transfers
- Introduced RTT delay to simulate real traffic
- Worst case (2-way) overhead measured ~5ms





Exploring the SDN Aspect cont. (3)

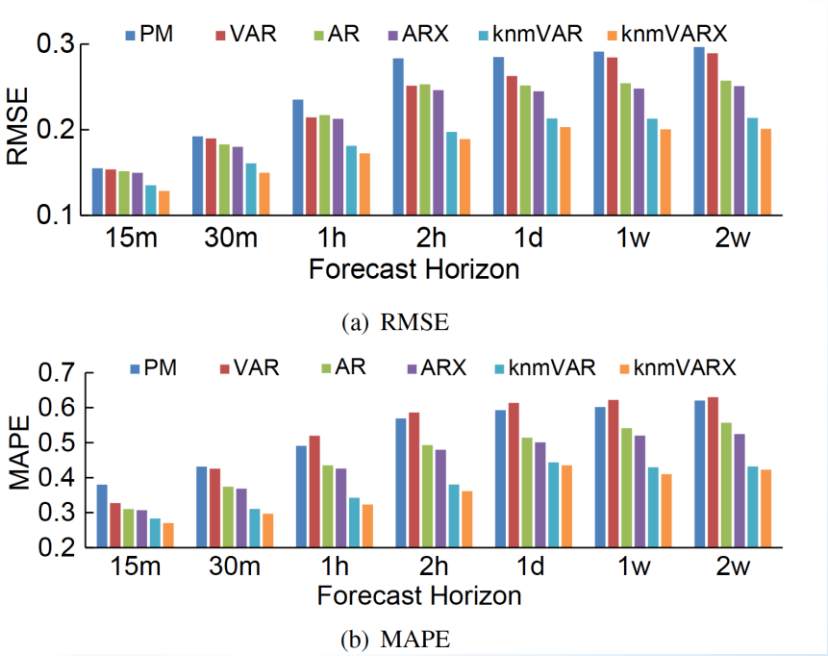
- Worst case scenario: wanted an experimental setup to get an idea of the overhead imposed on a flow by the SDN mechanisms
- Overhead impact diminishes with larger RTTs
  - But it's not negligible
  - The actual work on the payload has to be carefully planned to avoid excessive overhead
  - In the end, it's a cost vs. benefit question

Work in Progress

- Networking
  - Looking into what can be achieved through SDN, especially through SFC
- Algorithms
  - First use case: Smart Grid load forecasting

Spatial-Temporal Load Forecasting

- Smart Meter load data used to predict future grid load
- In comparison with various time-series baseline algorithms, best results come from the proposed knmVARX
- Early work was accepted in SmartGridComm2016



knmVARX: k nearest meter Vector Autoregressive with eXogenous weather input

Load Forecasting cont.

- First step towards streaming algorithm - need to establish a performance baseline
- Towards developing streaming version of algorithm to run on the wire:
  - knmVARX is based on global neighborhood statistics and requires data from all meters
  - Restrict to each network edge node or even to few higher-level nodes

Conclusions and Future Work

- Tested the feasibility of using SDN environment for on-the-wire processing
  - Depending on RTT, overhead does not seem to be prohibitive, even in the worst case
- Developed spatial-temporal baseline algorithm for Smart Grid load prediction
  - Better performance than known baseline algorithms
- Next Steps
  - Develop actual environment and deploy on faster hardware and software for further experimentation
  - Experiment with real payloads and algorithms, starting with streaming version of knmVARX

Thank You!

Questions?