TOWARD A NAME-BASED, DATA-CENTRIC PLATFORM FOR SCIENTIFIC DATA

LAN WANG (UNIVERSITY OF MEMPHIS)

LIXIA ZHANG (UCLA)

4/14/2020

A HUGE DATA EXAMPLE

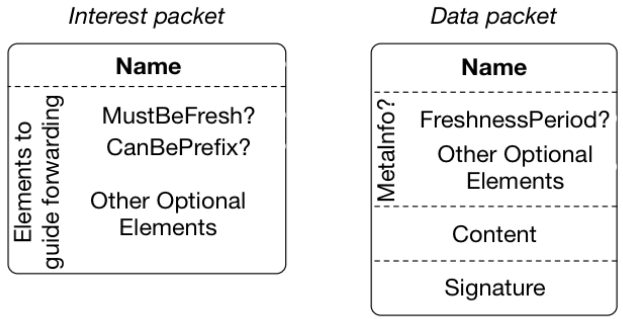
Consider a large study using data collected from thousands of participants

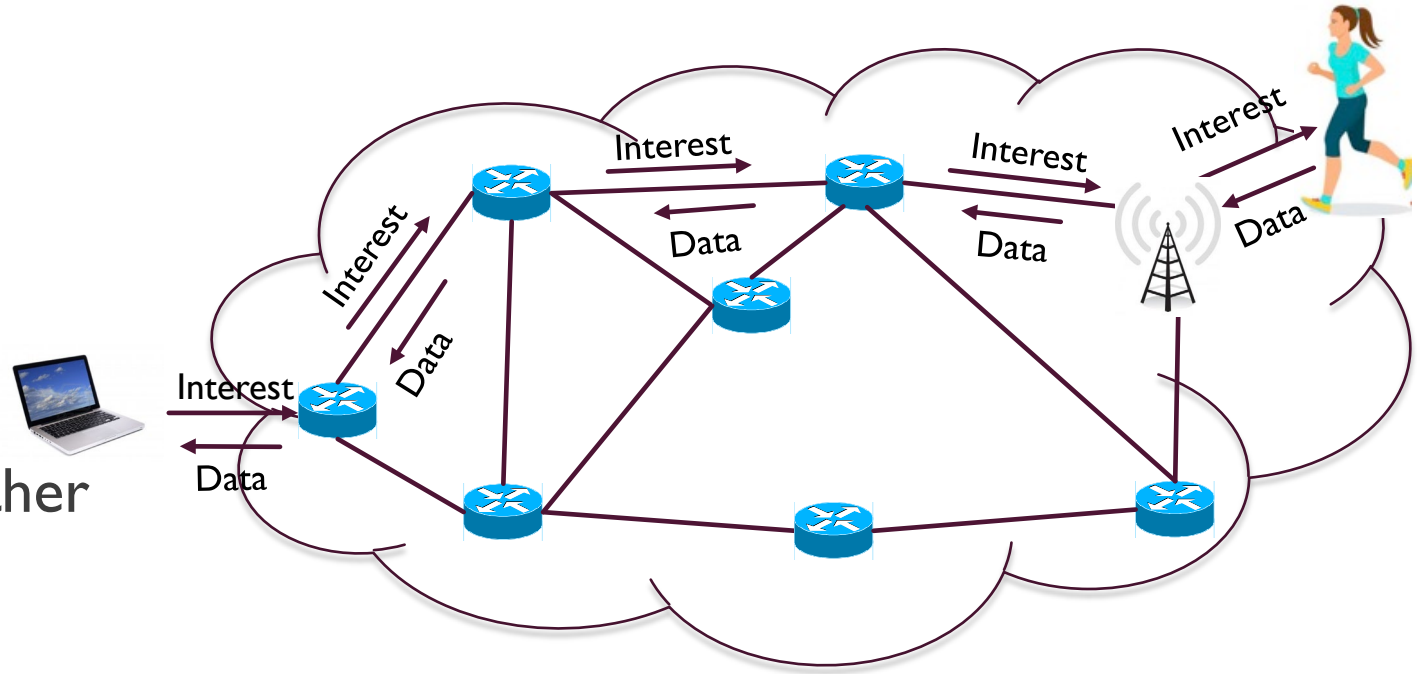
- How to collect, process, and distribute data in real-time?
- How to verify the integrity of computed results?
- How to ensure the data security and privacy of study participants?

Tokyo Marathon 2019 (right) and in 2020 (left). CREDIT: KYODO/REUTERS

NAMED DATA NETWORKING (NDN [1])

- **Name the data, not the container**
- Tell the network what you want (data name).
- Let the network find it for you.
- Verify data integrity and authenticity.
  - data signature
  - application-specific trust schema
- Intermediate nodes cache data for other users.

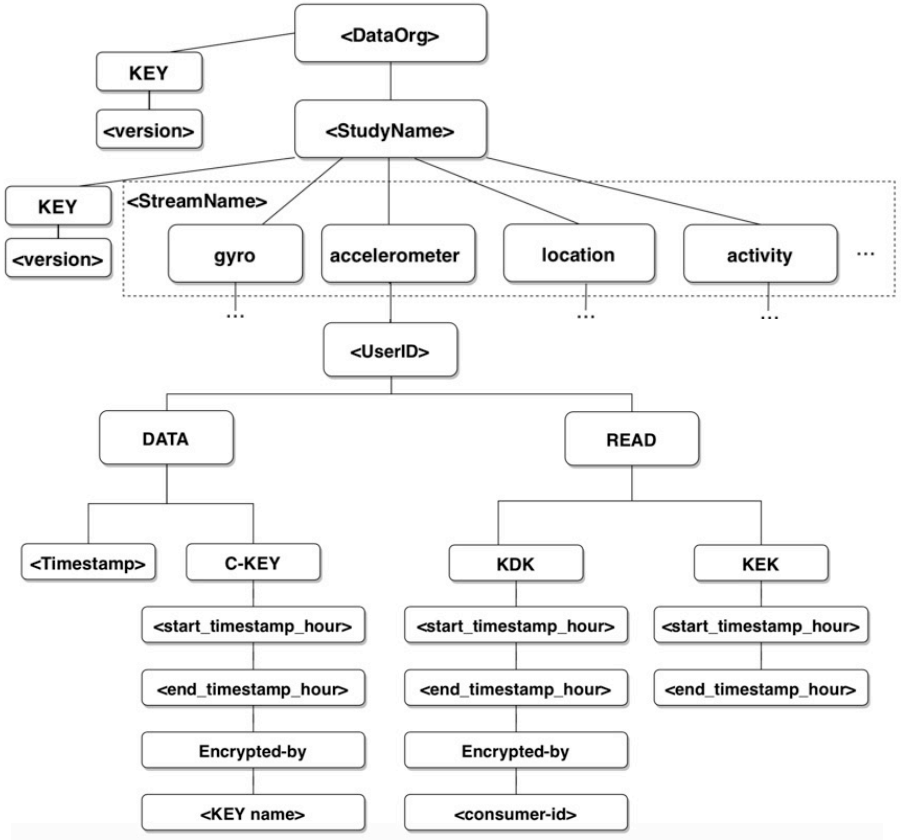




NDN packet forwarding

NAME-BASED DATA-CENTRIC PLATFORM

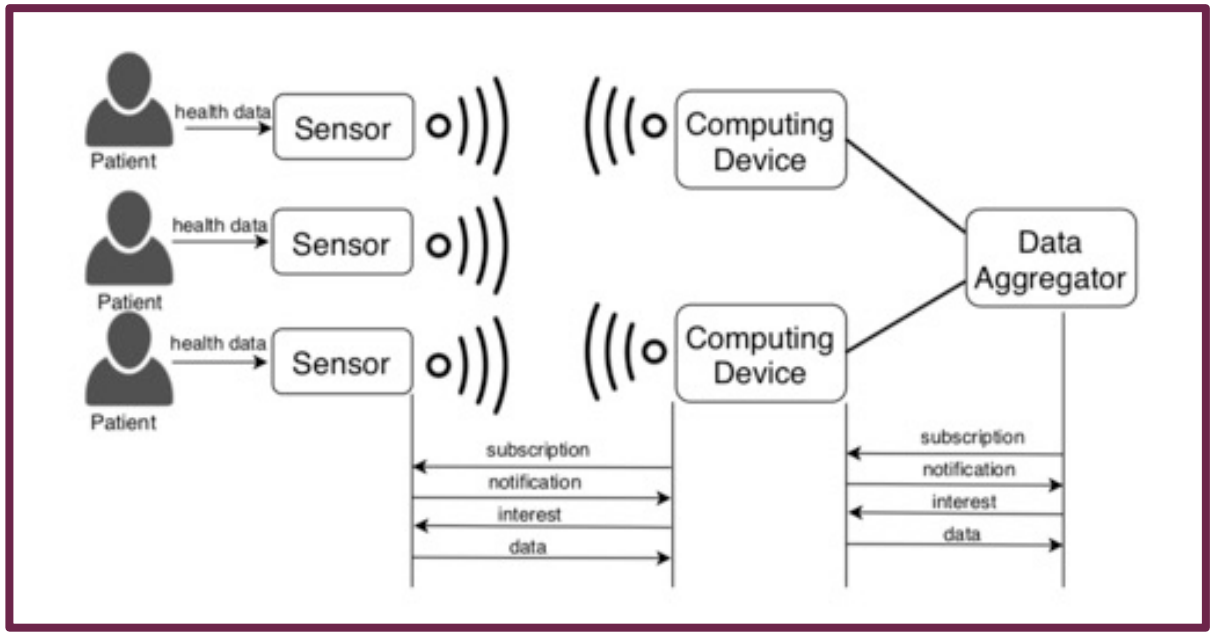
- Hierarchically named data
  - /<DataOrg>/<StudyName>/<StreamName>/<UserID>/DATA/<Timestamp>
    - Eg: /org/md2k/Tokyo19/gyro/alice/DATA/20000410053455
- Every piece of data contains a signature by the data producer.
- Individually named and secured data supports efficient data fetching, distributed processing, integrity, authenticity.



Sample Naming Scheme for a Study

DISTRIBUTED PROCESSING

- Data needs to be shared among sensors, computing devices, and end users.
- NDN Sync protocols [2] support multi-party data synchronization.
  - Use set reconciliation to sync the set of data names among group participants
- Pub-Sub API can be built on Sync [3].
  - A pub-sub system over NDN significantly simplifies the solution for distributed computing [4].
  - no centralized schedulers, DNS-based name translation, stateful load balancers, and heavy-weight transport protocols.



Distributed Computing of Health Data from Sensors through Pub-Sub over NDN

ENABLING DATA PROVENANCE

- How to verify data authenticity?
  - Data signing key indicates who produced the data.
  - NDN uses application-specific trust schemas [5] to automatically verify whether the owner of the key is authorized to produce the data.
- How to trace the series of computation and input data that led to a piece of data?
  - Use application naming scheme, data names, meta data

- What is new here?
  - Semantic naming enables systematic specification and automatic verification of the relationships between names (and the associated data).

Example: inferring activities from gyro and accelerator data

Input:
- gyro data: /org/md2k/mOral20/accelerator/alice/DATA/<timestamp>
- accelerator data: /org/md2k/mOral20/accelerator/alice/DATA/<timestamp>

Output:
- Inferred activity data: /org/md2k/mOral20/activity/alice/DATA/<starttime>-<endtime>/<compute-node>
- Meta data for activity data: /org/md2k/mOral20/activity-metadata/alice/DATA/<starttime>-<endtime>/<compute-node>
  - meta data contains (a) list of input data names, (b) inference algorithm, (c) parameters, (d) time of computation, …

NAME-BASED ACCESS CONTROL (NAC) [6]

- Define policies: who are given access to what dataset(s) with what restrictions.
  - Data users: /edu/memphis/lanwang
  - Datasets: /org/md2k/mOral20/gyro
  - Restrictions: data attributes and their ranges
- Enforce policies
  - Every piece of data is encrypted with a content key (C-KEY)
  - Encrypt and publish the C-KEY for only those users authorized to access the data

- How is this different from current access control (firewalls, SSL/TLS, …)?
  - End-to-end protection: data is encrypted both in transit and in storage.
  - Fine-grained: applied to data at every granularity, following the hierarchical name structure.

SUMMARY AND FUTURE WORK

- NAMING AND SECURING DATA DIRECTLY provides a foundation for
  - distributed processing
  - data provenance, confidentiality, automated fine-grained access control
- We are actively looking for collaborators to build a name-based data-centric software platform for large-scale scientific research over NDN.
  - Previous and ongoing NDN projects on supporting climate and HEP research [7, 8, 9]
  - More information about NDN software and testbed at www.named-data.net.

REFERENCES

1. L. Zhang, A. Afanasyev, J. Burke, V. Jacobson, K. Claffy, P. Crowley, C. Papadopoulos, L. Wang, and B. Zhang, "Named Data Networking," ACM SIGCOMM Computer Communication Review, July 2014.
2. T. Li, W. Shang, A. Afanasyev, L. Wang, and L. Zhang, "A Brief Introduction to NDN Dataset Synchronization (NDN Sync)," in IEEE MILCOM, 2018.
3. K. Nichols, "Lessons learned building a secure network measurement framework using basic ndn," in Proceedings of the 6th ACM Conference on Information-Centric Networking, 2019, pp. 112–122.
4. M. Król, S. Mastorakis, D. Oran, and D. Kutscher, "Compute first networking: Distributed computing meets ICN," in Proceedings of the 6th ACM Conference on Information-Centric Networking, 2019, pp. 67–77.
5. Y. Yu, A. Afanasyev, D. Clark, kc claffy, V. Jacobson, and L. Zhang, "Schematizing and automating trust in Named Data Networking," in *ACM ICN*, September 2015.
6. Z. Zhang, Y. Yu, S. K. Ramani, A. Afanasyev, and L. Zhang, "NAC: Automating access control via Named Data," in IEEE MILCOM, 2018.
7. C. Olschanowsky, S. Shannigrahi, and C. Papadopoulos, "Supporting Climate Research using Named Data Networking," in IEEE LANMAN, 2015.
8. S. Shannigrahi, C. Fan, and C. Papadopoulos, "Request Aggregation, Caching, and Forwarding Strategies for Improving Large Climate Data Distribution with NDN: A Case Study," in In Proceedings of ACM ICN, September 2017.
9. S. Shannigrahi, A. Barczuk, C. Papadopoulos, A. Sim, I. Monga, H. Newman, J. Wu, and E. Yeh, "Named Data Networking in Climate Research and HEP Applications," in 21st International Conference on Computing in High Energy and Nuclear Physics (CHEP2015), 2015.