WHAT'S IN A NAME?
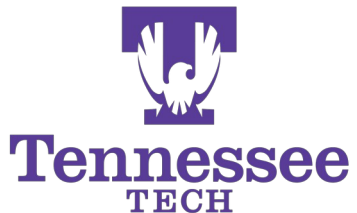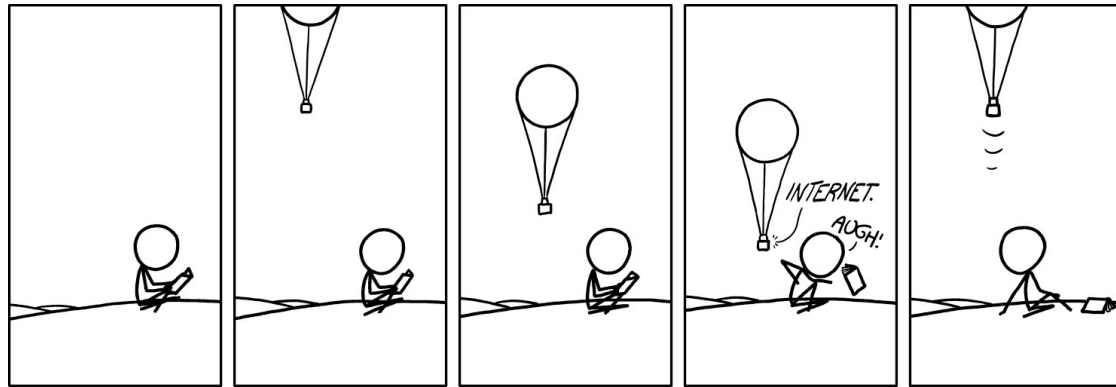# NAMING BIG SCIENCE DATA IN NAMED DATA NETWORKING

Susmit Shannigrahi, Chengyu Fan, Craig Partridge

sshannigrahi@tntech.edu
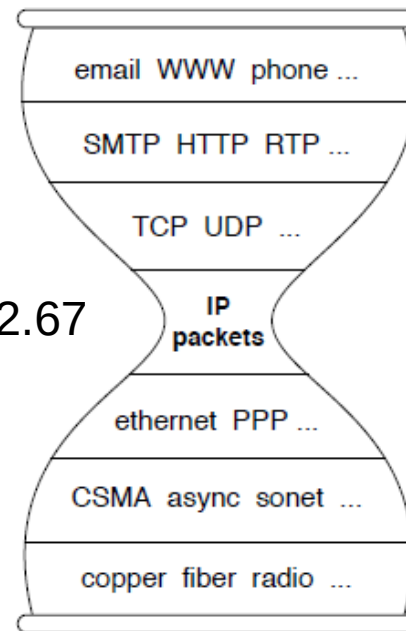
Names and the Network
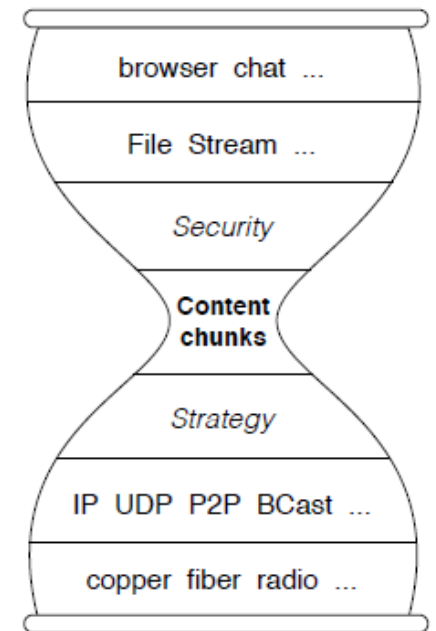
https://xkcd.com/1226/
https://1226.xkcd.com

151.101.192.67

Names in Named Data Networking
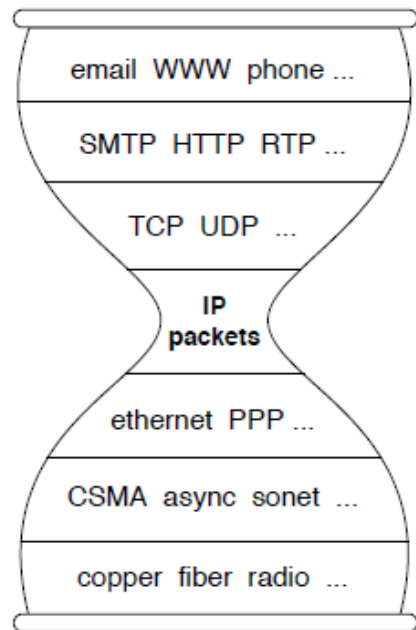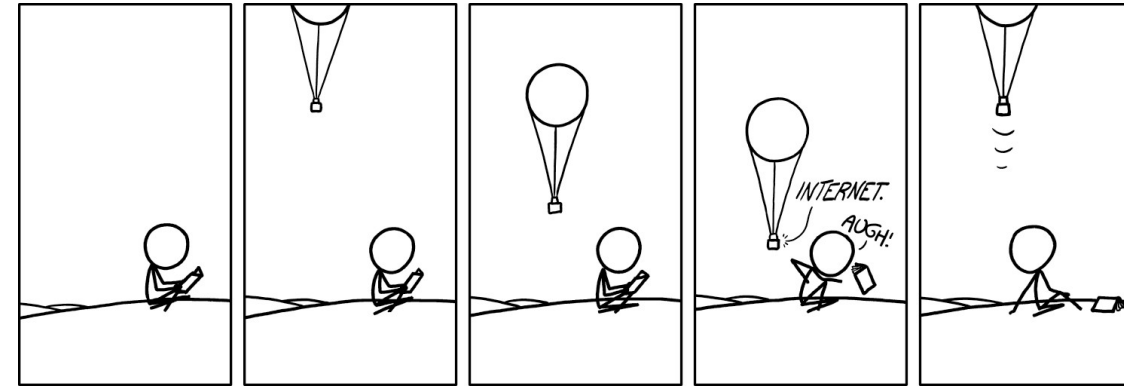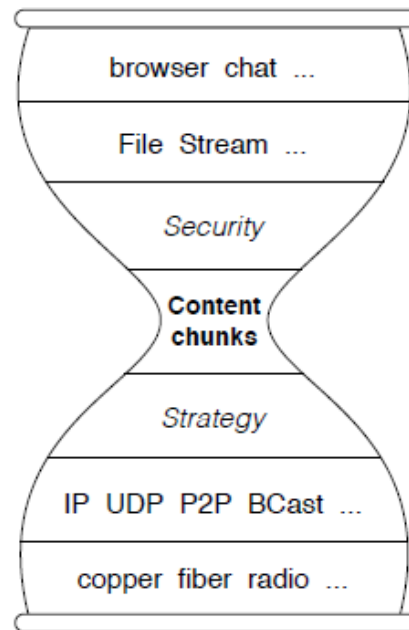
/xkcd/1226/
/1226/xkcd
/1227/xkcd
/1228/xkcd

FIB →
/xkcd/
/1226
/1227
/1228

Motivation

- Naming is *the* central construct for most NDN operations

- In NDN, naming is intertwined with network and application performance

- Lack of systematic study on NDN naming

Goal

- What names make sense in the context of scientific applications?
  - Can we create them from existing names?

- What are the trade-offs of such names?

- A number of scientific communities can benefit from NDN
  - Can we come up with a list of naming recommendations for future applications?

Outline

- Motivation

- **Existing Names and Translating them in NDN**

- How does naming affect NDN operations?

- Naming recommendations

Existing Names in Scientific Communities

- Follows community driven hierarchical naming schemes
  - Community driven
  - Strict enforcement is often lacking

- Follows directory structure for data management
  - Files need to be identified as part of a larger collection

- Names have domain specific information
  - needed for applications and users

A Climate File Name

http://esgf-data1.diasjp.net/thredds/fileServer/esg_dataroot/
cmip5/output1/MIROC/MIROC4h/historical/6hr/atmos/6hrPlev/
r1i1p1/v20110926/psl/psl_6hrPlev_MIROC4h_historical_r1i1p1_
1984010100-1984123118.nc

/CMIP5/output/MIROC/MIROC4h
/historical/6hr/atmos/psl/r1i1p1/1950010100-1950123118/

NDN Name

8

A Root File Name (Physics)

root://eospublic.cern.ch//eos/opendata/cms/MonteCarlo
Upgrade/RunIIFall18wmLHEGS/TTToHadronic_TuneCP5_
13TeV-powheg-pythia8/GEN-SIM/IdealConditions_102X_
upgrade2018_design_v9_ext1-v2/260000/
4FF18384-7B96-9E4A-B611-AA5D176A4156.root

/cms/MonteCarloUpgrade/RunIIFall18wm
LHEGS/TTToHadronic_TuneCP5_13TeV-powheg-pythia8
/GEN-SIM/IdealConditions_102X_upgrade2018_design_v9_
ext1-v2/260000/<UUID.root>

NDN Name

A Genome Name

Original Name

https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR5139424

/BIOLOGY/SRA/9605/9606/NaN/RNA-Seq/ILLUMINA/
TRANSCRIPTOMIC/PAIRED/Kidney/PRJNA359795/
SRP095950/SRX2458154/SRR5139424/SRR5139424_1

NDN Name

The takeaways

- The concept of hierarchy exists in scientific communities

  - Some communities have already named their data

  - Others can convert their names into these *community defined* hierarchies

  - Good news for NDN!

Outline

- Motivation

- Existing Names and Translating them in NDN

- **How does naming affect NDN operations?**

- Naming recommendations

NDN Names – Unresolved Questions

- Possible to create hierarchical and semantically meaningful names

- How much information should be included?

- In which order?

- How should the network handle them?

- What names should the applications use?

- Which prefixes should be announced into the network?

Expressive vs Shorter names

- A design continuum rather than exact lengths

- Expressive names carry more information
  - Encodes more domain specific information
  - Provide better context for both the network and applications
  - e.g., probabilistic caching → cache any content that has "cmip5" in the name
  - Forwarding → More granular namespaces

- Shorter names are better for reducing in-network states
  - Less useful for providing context
  - Less overhead (generally)
  - Smaller in-network state

Expressive vs Shorter names

/genus/species/subspecies

	Expressive Names	Shorter Names
Caching	(=) Popularity Based (+) Better context for probabilistic/Predictive/Context aware caching	(=) Popularity Based
Forwarding	(+) Fine grained forwarding (-) Can be slower (-) Overhead is higher	(+) Less State (+) Potentially faster (-) Coarse granularity
Routing	(+) Better route choices (-) More announcements	(+) Fewer announcements (-) Producer has less control over forwarding

Expressive vs Shorter names – Data Structures

/genus/species/subspecies/…/

	Expressive Names	Shorter Names
Interest/Data Packets	(-) Larger, less space for payloads	(=) Popularity Based
FIB/RIB	(-) Larger	(+) Smaller
Measurement Tables	(-) Larger	(+) Smaller

Expressive vs Shorter names – Application Layer Functionality

/genus/species1/subspecies
/genus/species2/subspecies

	Expressive Names	Shorter Names
Publication	(+) Needed for partial replication (-) Can use forwarding hints	(-) Difficult to support partial replication
Data Discovery	(+) More context is better for discovery and search	(-) Less context
Privacy	(-) Exposes more information	(+) Exposes less information

Decision Time

- Expressive names are better suited for both network operations and applications

- The trade-offs are: smaller payload and larger in-network state

- Which one should the scientific communities prefer?
  - Depends on the application and the network operator
  - We are only pointing out the trade-offs

Minimum Usable Data Units (MDU)

- In science communities, a set of files are often used together
  - Single files are useless

- Makes sense to treat them as a single entity
  - the Interest and Data packets should be handled consistently (e.g., same upstream and downstream, same strategy)
  - Cached together and subsequent requests sent past it

- **We are not proposing prefetching**
  - A hint to the network layer, not a requirement
  - Can be implemented as a strategy

```
/Trichosporon/asahii/1.1/ht2
/Trichosporon/asahii/1.2/ht2
/Trichosporon/asahii/1.3/ht2
/Trichosporon/asahii/1.4/ht2
/Trichosporon/asahii/1.5/ht2
/Trichosporon/asahii/1.6/ht2
/Trichosporon/asahii/1.7/ht2
/Trichosporon/asahii/1.8/ht2
/Trichosporon/asahii/1/fa
/Trichosporon/asahii/1/gff3
/Trichosporon/asahii/1/gtf
/Trichosporon/asahii/1/meta.json
/Trichosporon/asahii/1/Splice_sites
```

Outline

- Motivation

- Existing Names and Translating them in NDN

- How does naming affect NDN operations?

- Naming recommendations

Naming Recommendations for Applications

- Constructed using well-defined, hierarchical, human-readable, and semantically meaningful components
  - Provides more context to the users, applications, and network operators

- Longer names are more useful
  - Trade-off is space, can use linked objects for both shorter and longer names

- Don't use location information as part of the name

- Allow the names to be identified as part of the MDU

- For larger namespaces, use forwarding hints

- Put any sensitive information towards the end and encrypt it

Naming Recommendations for Science Network Operators

- Shorter names help with forwarding speed (generally), expressive names help with everything else (caching, forwarding)

- Data names should be location agnostic
  - */cmip5/csu* is easier to handle in NDN than */csu/cmip5*

  - Unless you need to know the location
    - e.g., Load-balancing between multiple sources

- Cache and forward MDUs together for better efficiency

Conclusions

- Naming is the central construct of NDN

- Efficient naming can align named data with application requirements and network characteristics

- For science data, we define the concept of MDU
  - handle a set of files together
  - benefits both the application and the network

- We provide a list of recommendations
  - Up to the community and the operators to implement them