RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

Computing in the Continuum: Harnessing Pervasive Data Ecosystems

Manish Parashar, Ph.D.

Director, Rutgers Discovery Informatics Institute RDI^2

Distinguished Professor, Department of Computer Science

Moustafa AbdelBaky, Daniel Balouek-Thomert, Eduard Renart, Mengsong Zou, Ali Reza Zamani, and Javier Diaz-Montes

In collaboration with Ericsson and Cardiff University

International Workshop on Urban Data Science

Rochester Institute of Technology, Rochester, NY

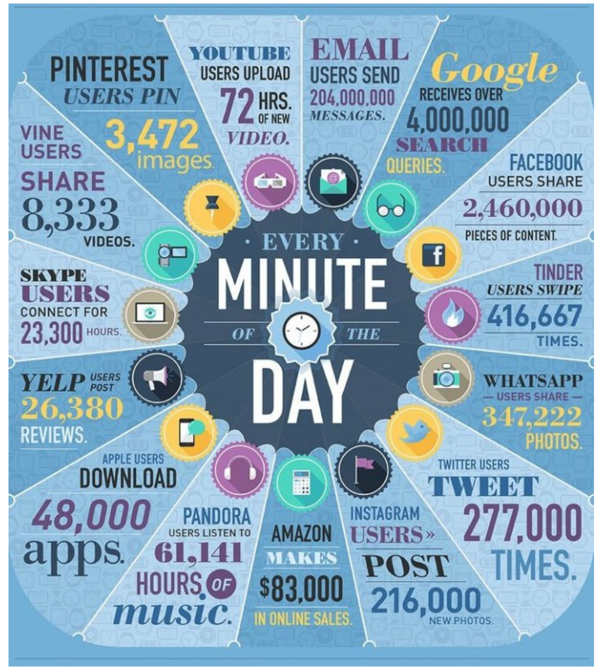
June 28, 2017

Outline

- The emerging data/compute reality
- Computing in the Continuum
- Research challenges (and some ongoing research)
- Conclusion

The Era of Big Data and Extreme Compute

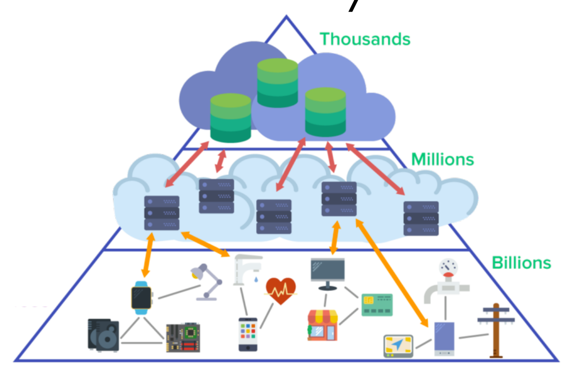
- *Data and computing data are pervasive*
- Extreme scales; extreme data volumes and rates
- Novel paradigms: cloud services everywhere, cloud/fog/edge, in-transit, SDN/NFV, IoT, …
- New technologies: accelerators, storage, communication, …
- New concerns: correctness, energy, fault tolerance, etc.



Entire printed collection of the US library of Congress is 10 Terabytes

1 Exabyte is 100,000 US Libraries of Congress

One BILLION BILLION operations per second by 2021!

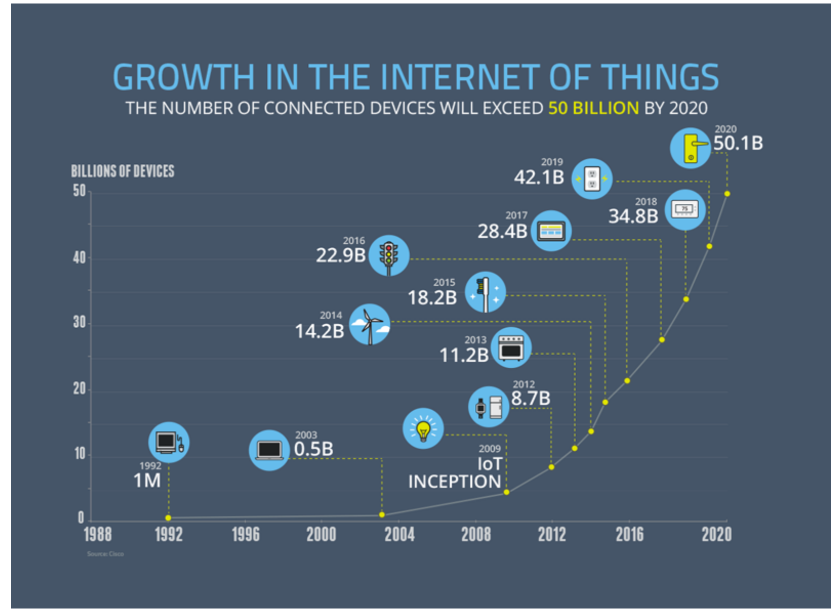


Society Transformed by Data & Compute

- New "data rich" reality in virtually every domain
- Increasing need for data-driven online processing…

The New Reality – Dynamic, Data Driven!

- Unprecedented instrumentation and the extraordinary growth of digital data sources have led to an explosion of data
  - Large size, heterogeneous in nature and geographic location
  - New opportunities; new challenges

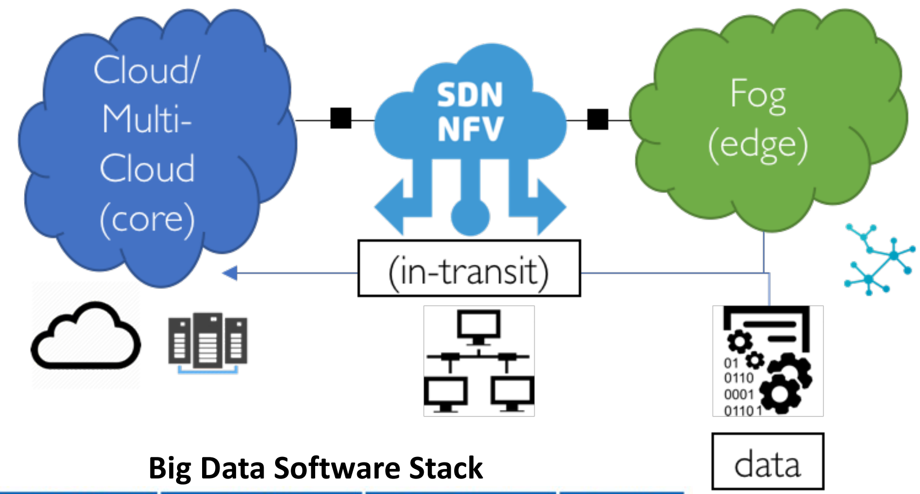


50 BILLION IoT DEVICES BY 2020

- Large volumes of data that need to be processed by complex application workflows in a timely manner
  - Use cases span precision medicine, smart infrastructure, instrumented oilfields, disaster management, etc.
- Need to rethink existing paradigms and practices

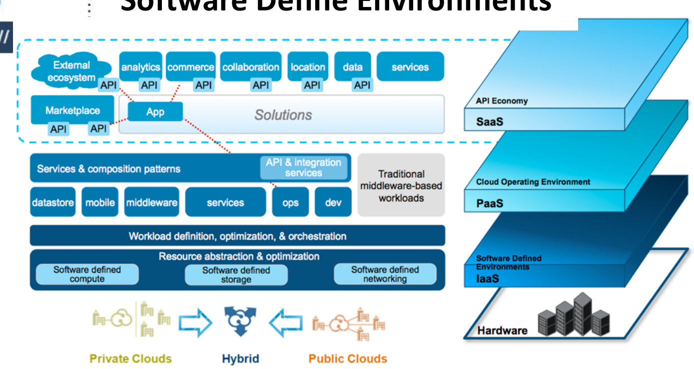
An Evolving Services Landscape

- Cloud
  - Hosted in data centers at the core
  - Relatively inexpensive; seemingly infinite
  - Far from the data; data access is expensive
- Fog/Edge
  - Computation/storage is limited and expensive
  - Closer to the data; lower latencies
  - Limited and unreliable connectivity
- In-Transit
  - Distributed along the data path
  - Limited, but can be effective
  - Intermediate latency
  - Fewer guarantees

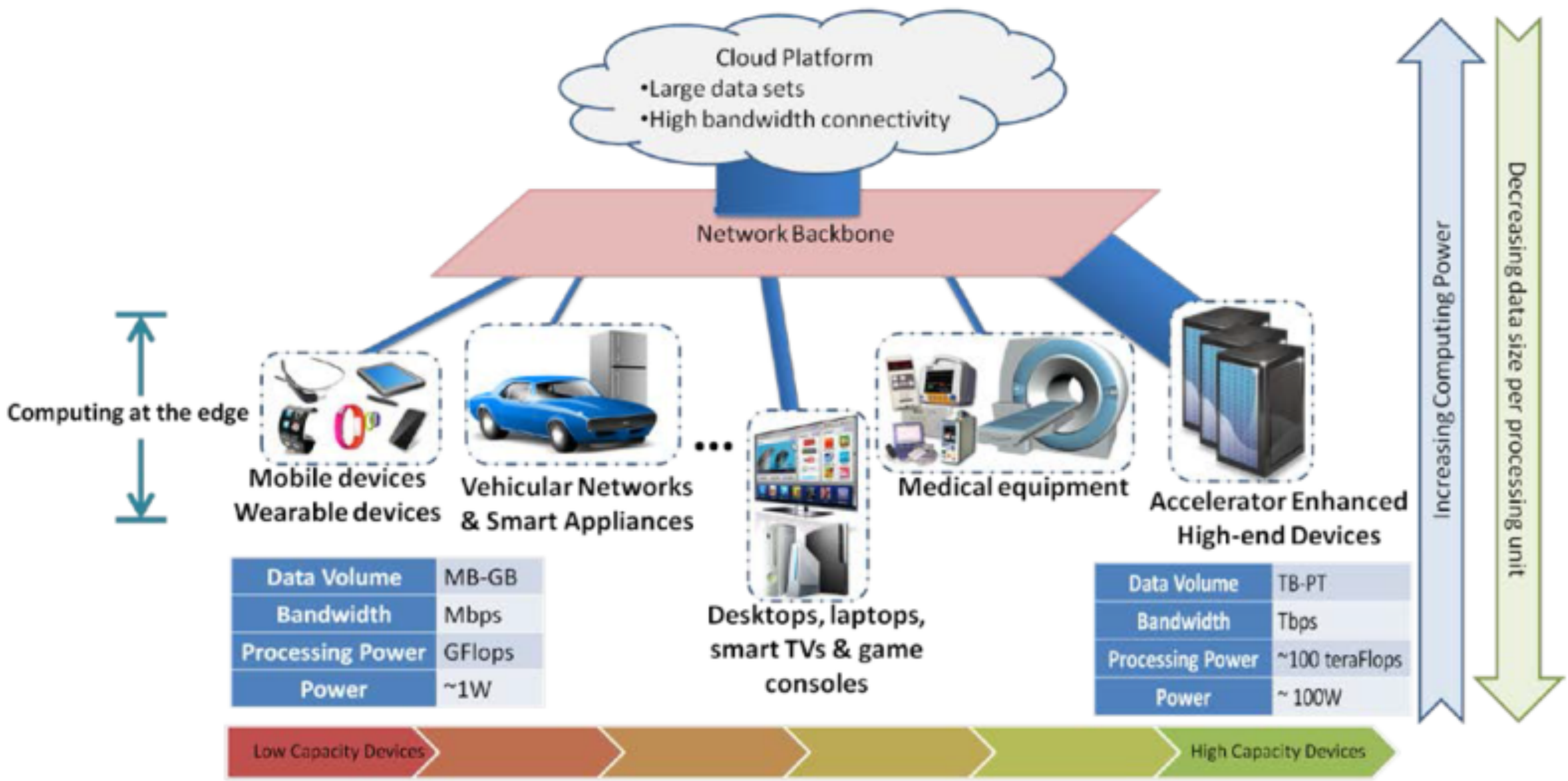


Big Data Software Stack

Software Define Environments

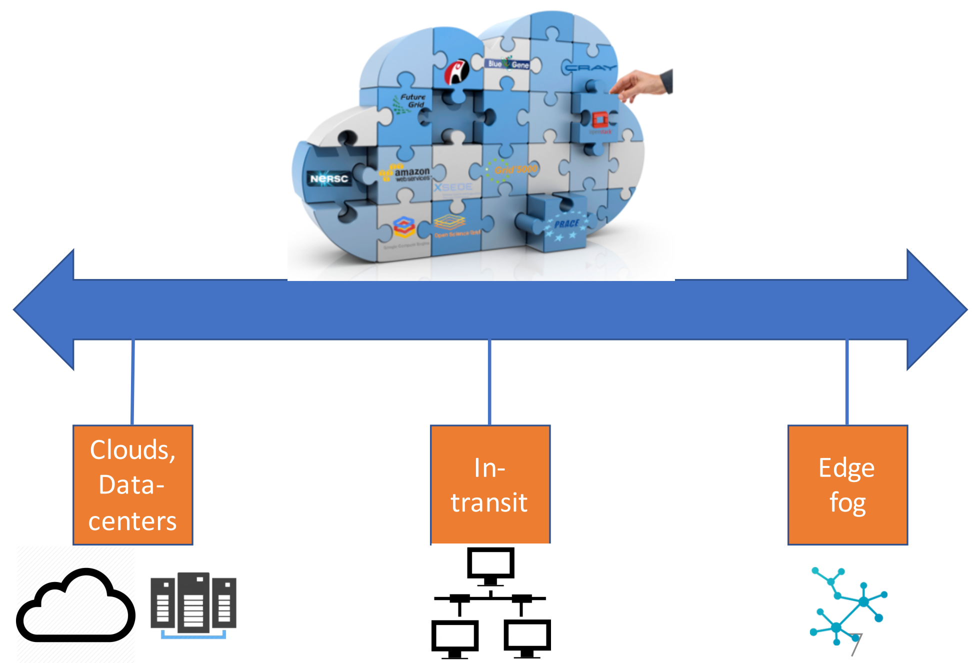


The New Reality – Dynamic, Data Driven!



Computing in the Continuum

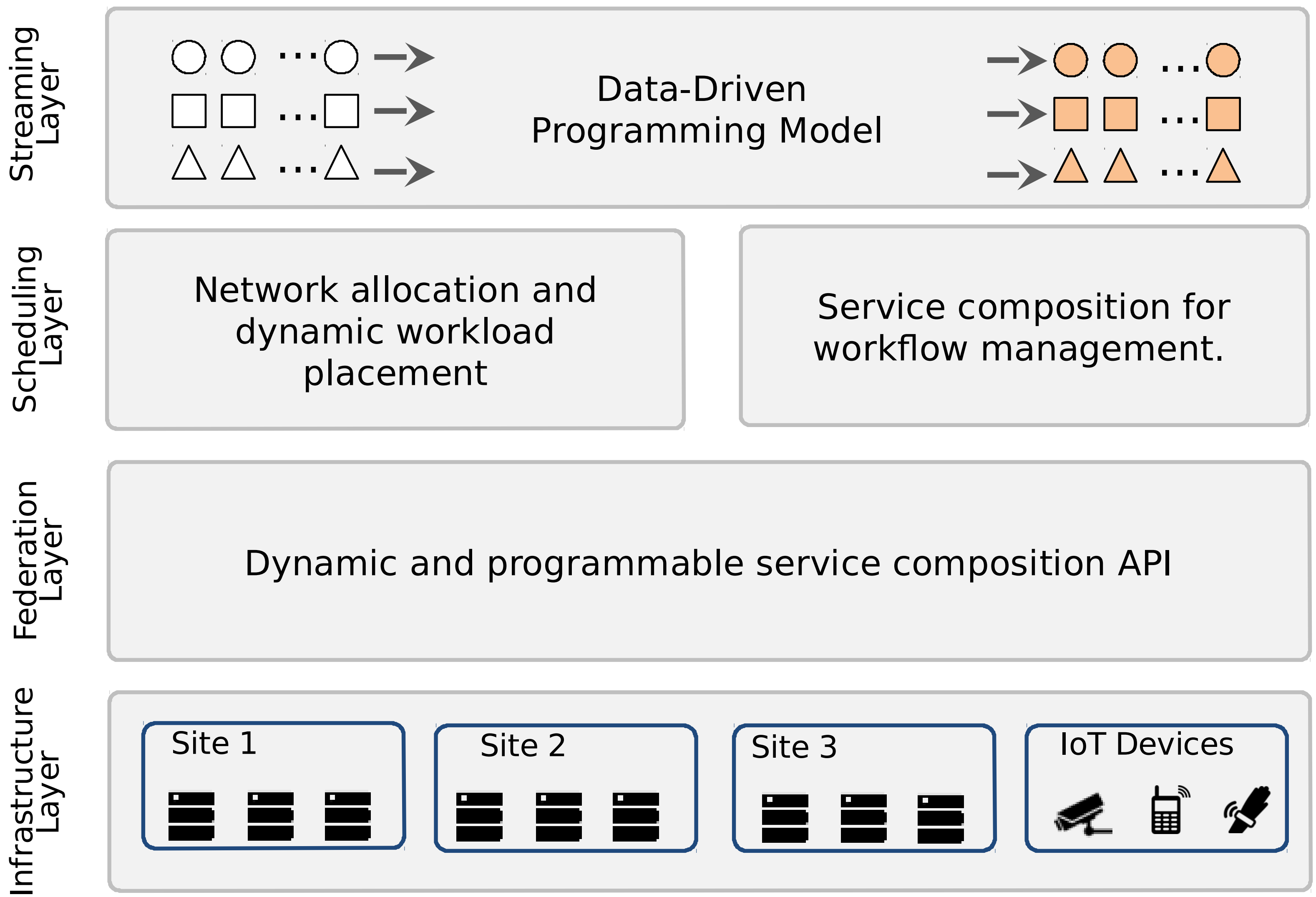
- Leverage resources and services at the logical extreme of the network and along the data path to increase the value of the data while potentially reducing costs
- Exploit the rich ecosystem of data and computation resources at the edge so that **data is not moved**
- Identify the **high levels of concurrency** that is pervasive throughout the ecosystem as the key to realizing scalable data-centric applications



Research Challenges

- How to drive computation through data
  - Express application behavior based on available data and its content
- How to accommodate uncertainties in data and computation
  - Move away from precise to approximate computing
- How to build applications and manage workflows so that they adapt to increase their value
- How to continuously optimize execution in a dynamic data-driven environment
  - How to discover and aggregate services (data, resources, …) that fit the current requirements
- How to incorporate market models, social/trust models, and utility models
- …….

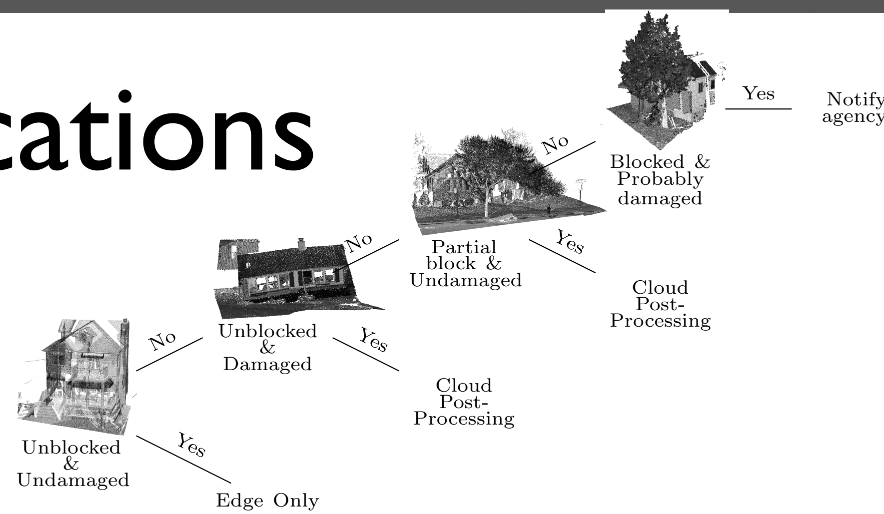
Computing in the Continuum – Some Research Efforts at RDI^2



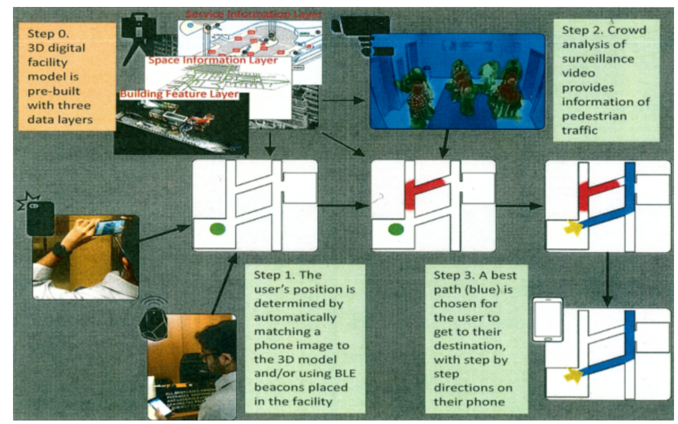
RUTGERS DISCOVERY INFORMATICS INSTITUTE

Some Driving Applications

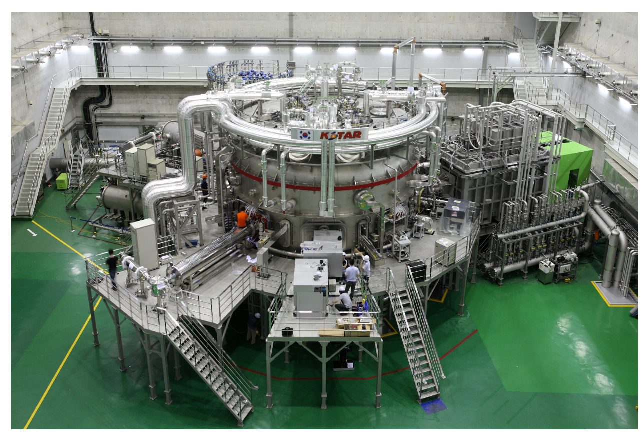
1. Data-driven emergency response to structural damage of civil infrastructure



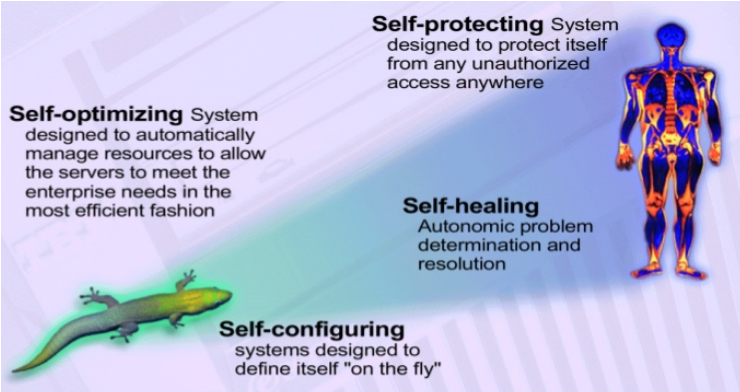
2. Enhancing mobility for autistic individuals using urban sensing



3. Scientific experiment management



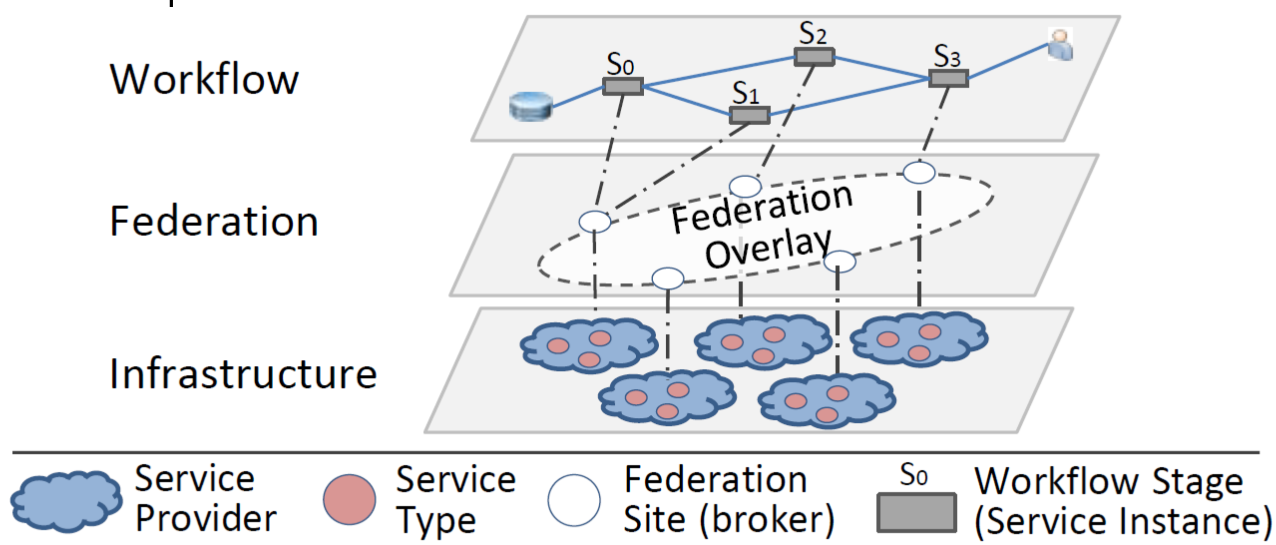
Autonomic Management



- Optimize resource provisioning and workload allocation to meet objectives and constraints set by users, applications, and/or resource providers
- Create models to translate resource/service capabilities and availabilities into application-level utilities (e.g., throughput, performance, etc.)
- Combine predictive and reactive approaches to improve decisions
- Quantify errors and uncertainties to offer confidence levels
  - How much error can I tolerate to maintain certain QoS?

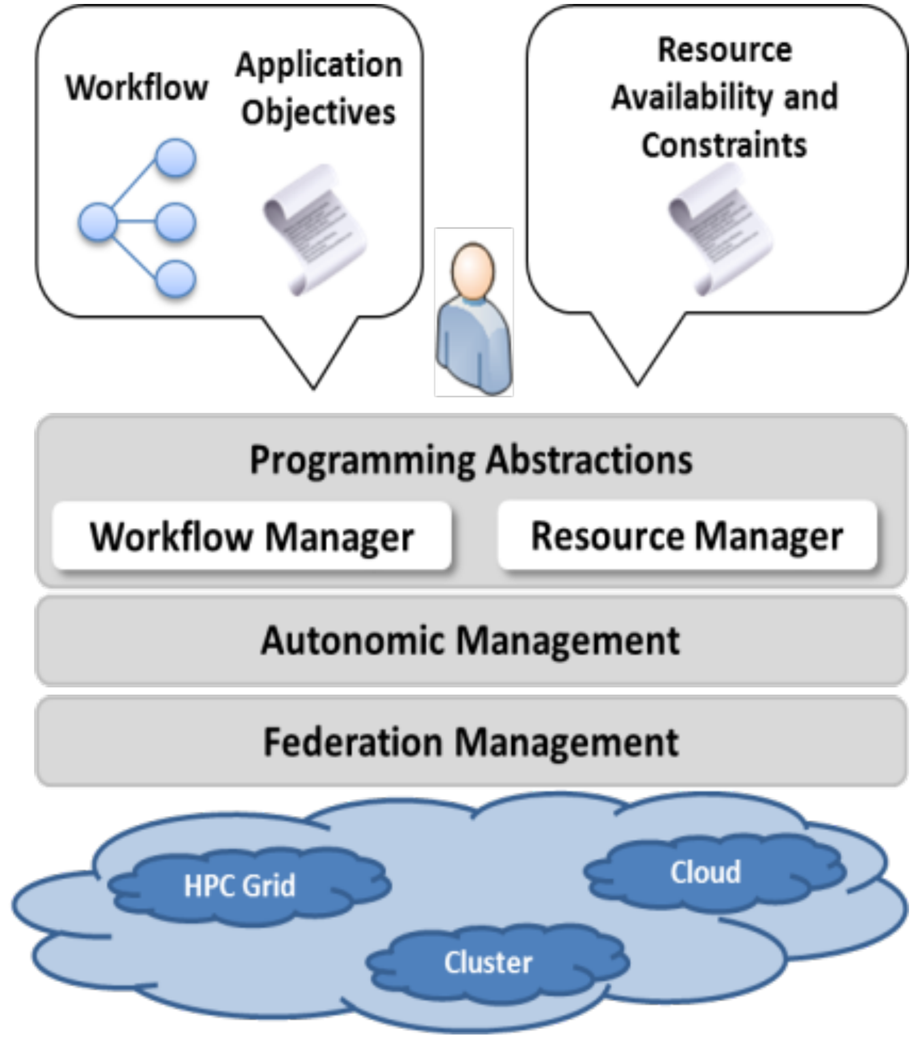
Dynamic and data-driven application workflows

- Continuous workflow self-management/optimization
  - Dynamically discover and compose services (data, software, resources, etc.) / synthesize custom application workflows and execute platforms on-demand, across the computing continuum
  - Evolve dynamically (and opportunistically) based on service availability, failures, location of data producer or data consumer
- New data driven programming models/abstractions required to support aspects such as opportunistic elasticity, uncertainty, approximation, continuous optimization
- Utility/economic models – applications centric cost/benefit analysis



Distributed Software-defined Environments

- Combine cloud service abstractions with ideas from software-defined environments
- Create a nimble and programmable ecosystem that autonomically evolves over time, adapting to:
  - Changes in the infrastructure
  - Application requirements
- Independent control over application and resources
- Enable efficient processing to transform data into actionable knowledge that drive critical decision making
  - Allocate computational power close to digital data sources
  - Process data in-situ and/or in-transit



M. Abdelbaky, J. Diaz and M. Parashar, "Enabling Distributed Software-Defined Environments Using Dynamic Infrastructure Service Composition," CCGrid'2017

Distributed Software-Defined Environments

Applications & Workflows

- Description and dependency
- Workload behavior
- QoS objectives (e.g., deadline, budget)

- varies at runtime -

Autonomic Management

- Identify utility of composition
- Ensure objectives are always met
- Adapt & reconfigure services

- on the fly -

User/Resource Provider

Define composition programmatically using rules and/or constraints

- Availability
- Application requirements (e.g., type of resources, memory, I/O rate)
- Cost
- Location
- Access policy

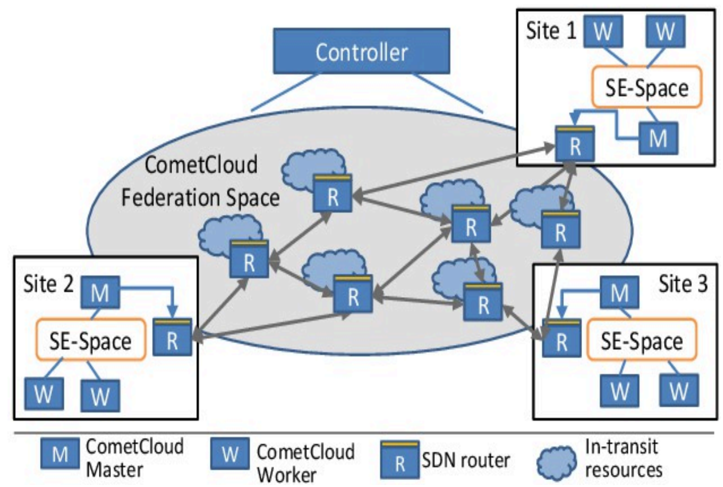
- varies at runtime -

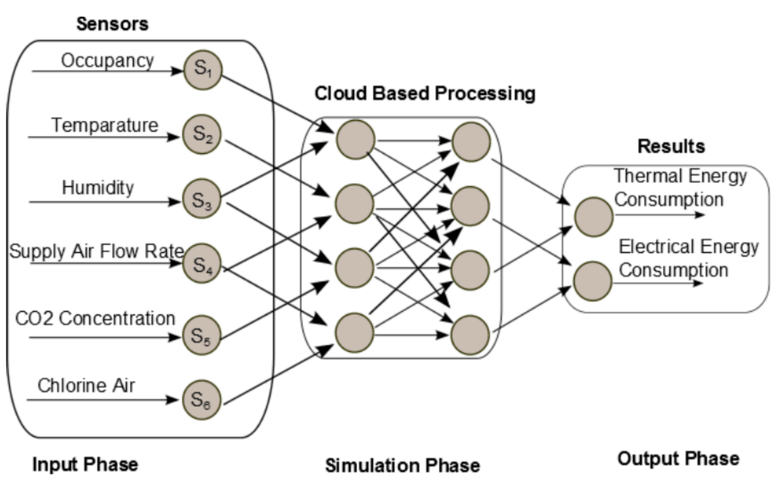
Exposed as a service to applications & workflows

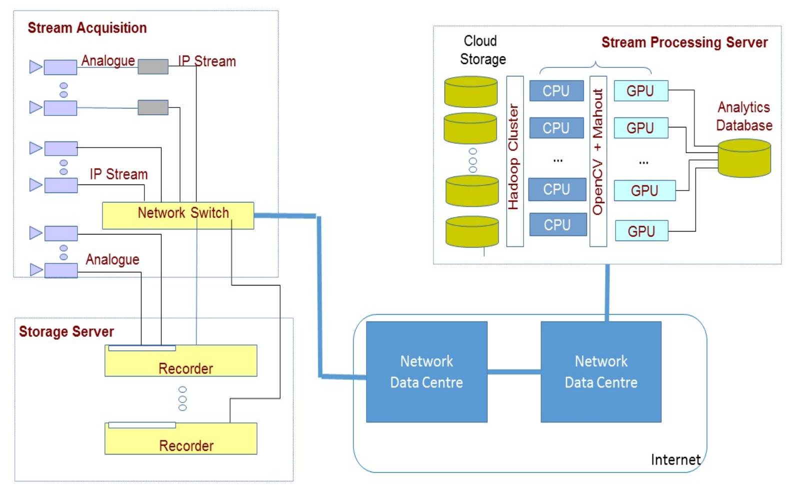
Synthesize a space-time ecosystem

In-transit Computing (with O. Rana, I. Petri, Cardiff)

- Use computational capabilities along the path
  - Leverage excess (idle) resources within the network infrastructure
  - Enable general purpose computation in the network data center
- Extend SDN controllers to manage resources at network data centers
  - Resource provisioning/allocation for in-transit processing
  - Explore approximations in-transit







A. Zamani, M. Zou, J. Diaz-Montes, I. Petri, O. F. Rana, A. Anjum and M. Parashar, "Deadline Constrained Video Analysis via In-Transit Computational Environments", IEEE Transactions on Services Computing. 2017.

Data-driven Programming

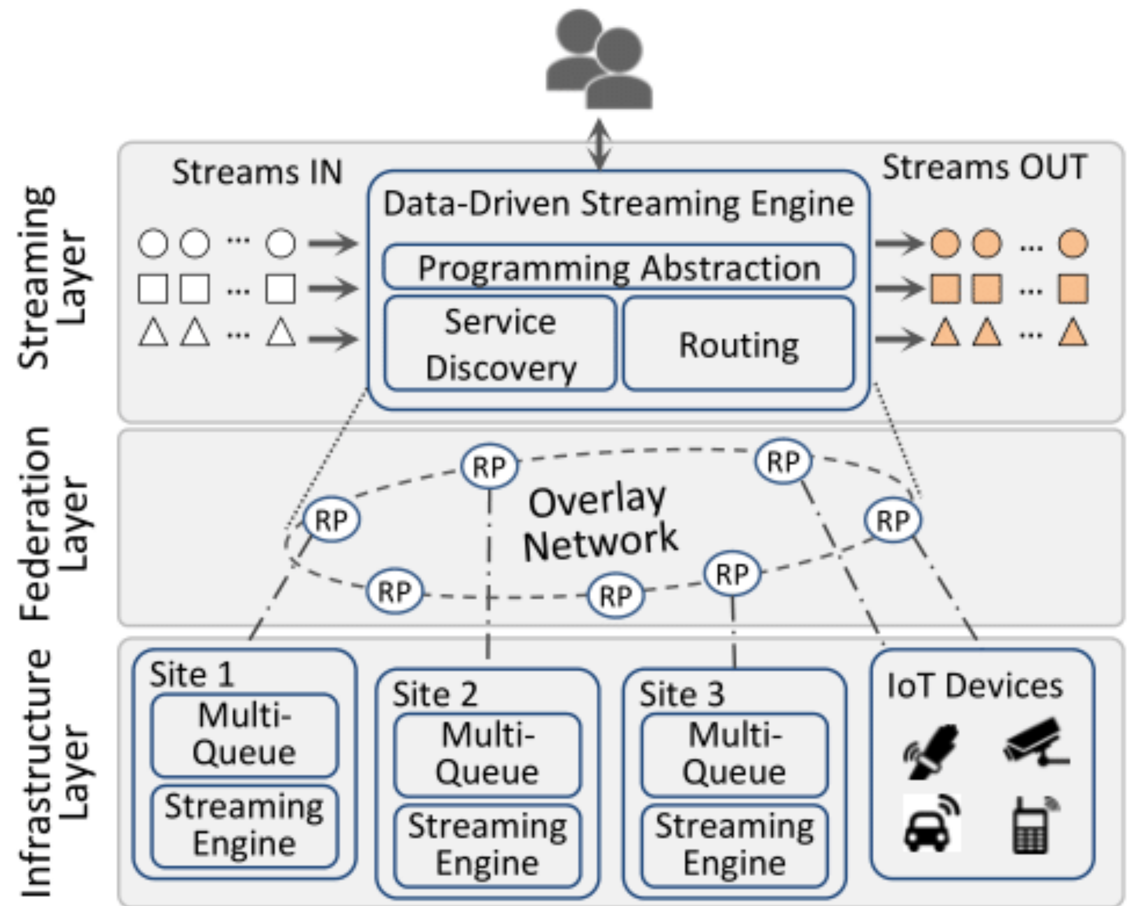
Data-drive Stream Processing: Location aware; Content aware; Quality/Uncertainty aware

Major stream processing engines do not support data-driven, location-aware, resource-aware operators

	Apache Kafka	Google Pub/Sub	Amazon Firehose
Data Organization	Topic name	Topic name	Stream name
Producer/Consumer Discovery	No	No	No
Query data by content	No	No	No
Location aware	No	No	No
Resource aware	No	No	No
Data consumption	One-to-one	One-to-many	One-to-many
Data subscription	Topic	Topic	Stream
Can be Deployed	Core or Edge	Core	Core

Data-driven Programming

- Data-drive Stream Processing in the Continuum: *Location, Content, Quality/Uncertainty* aware
- Content-based, event-driven specification of which topology to execute; where to execute it
- Builds on the *Associative Rendezvous* paradigm for content-based decoupled interactions with programmable reactive behaviors
- Implementation based on Strom and Kafka



E. Renart, J. Diaz-Montes, and M. Parashar, Data-driven stream processing at the edge. In IEEE International Conference on Fog and Edge Computing (ICFEC), May 2017

Some Related Projects

- ExoGENI
  - A framework that orchestrates a federation of independent cloud sites located across the US and circuit providers (e.g., NLR and Internet2) through their native IaaS API interfaces, and links them to other GENI tools and resources
- FELIX
  - A federation framework that allow users to build their own virtual slices using resources of remote Internet facilities
- CloudLab/Chameleon
  - Experimental meta-cloud environments that provide bare-metal/vm access and control across multiple university sites
- Atos
  - Multi-cloud service integration
- DITAS
  - Simplified logistics/data-management for cloud and edge environments

Summary

- Computing in the continuum
  - Leverage resources and services at the logical extremes, along the data path, in the core…
- Software-defined ecosystems
  - Living workflows
  - Software defined systems
  - Autonomic management / continuous optimization
- Many challenges at multiple layers
  - Application formulation, programming systems, middleware services, standardization & interoperability, autonomic engines, etc.

Thank You!

M. AbdelBaky, D. Balouek-Thomert, E. Renart, A. Zamani, M. Zou, M. Parashar

Rutgers Discovery Informatics Institute (RDI^2)

Rutgers, The State University of New Jersey

Email: parashar@rutgers.edu

WWW: http://parashar.rutgers.edu/

CometCloud: http://cometcloud.org