RESEARCH ARTICLE OPEN ACCESS

# Big Data: Data Driven Design philosophy for Solving Hard Optimization Problem

Sridevi Malipatil [1], Vani. N [2]
Department of Computer Science and Engineering
RYMEC (Rao Bahadur Y Mahabaleswarappa Engineering College)
Ballari - Karnataka

## ABSTRACT

Big data brings the era for fourth paradigm in science discovery through data-driven design philosophy for finding hard optimization problem. This paradigm gives a new approach for designing future Internet. The existing system does not support new applications and no efficient resource utilization. The proposed approach solves these issues by supporting new applications with efficient resource utilization technique. The data driven design upgrades in architecture of network, services, and applications, and performs design of future Internet architecture, communication models, and resource management mechanisms.

***Keywords:-***Hadoop, SDN, Big Data, ICN

## I. INTRODUCTION

Big-data refers to large scale information management. Big data analytics is the process of analysing large data sets containing different data types. Big Data analytics can be employed to analyse huge data in internet which performs transactions in online banking and identify security threats and suspicious activities, and to correlate multiple sources of information into a coherent view.

The latest trends in design of future internet requires network expansion, resource provisioning have introduced new challenges in network science and engineering, summarized as follows.

- **Availability**: The availability of both network infrastructure and services is essential as the scale and variety of network applications and change in network speed performance. For example, video conference will represent 86 percent of global Internet traffic, but the Internet was designed for data file transfers, thus having difficulties in providing highly available video services.

- **Efficiency**: The future Internet is to deliver customized information in an increasingly effective manner. Social network applications help personalize content and services, make the content consumption highly selective, and require the network to be agile in content delivery through Facebook 1.3 billion. Internet data center (IDC) investment is increasing day to day and networks have become more software programmable to make efficient usage of servers and resources and to decoupled control and data plane protocols (e.g., Open Flow). Such software defined networks (SDNs) enable large, real-time, and automatic network resource scheduling and control.

- **Evolvability:** The future Internet is an ever evolving composite of devices, services. The Internet architecture is an open-ended process as infrastructures and applications change steadily. Since there is no access to an accurate prediction of future network services, any proposed future Internet architecture must be able to work over time.

- **Computational Intelligence:** the computational complexity is replaced by state complexity. Instead of maintaining complex network states in all management and control device, the design of future Internet computes the network status and applies policies at runtime programmatically on network devices with simplified data plane.
- **Data Intelligence**: Data obtained at scale bring both entropy and intelligence. Comprehensive data outweigh sampled data, correlation is used rather than causality from data analysis, and the efficiency of the network correlation discovery is emphasized over accuracy.

- **Effectiveness Intelligence:** Big data in network science has enabled the methodology shifted from mathematical approaches to effectively solvable, data-driven approaches for addressing issues in network resource management.

## II. OBJECTIVES

The objective of Data driven design is to make statements about collective behavior. Most of analysis applications involve privacy; protection of privacy is a main challenge. In order to design new Big Data Analytics services, there is a need for a new framework where actors can exchange data from different sources and provides privacy. Next challenge is to design solutions that will not reveal business critical information when disseminated. Identify techniques to enhance security in big data frameworks (e.g., data tagging approaches, Hadoop).

## III. PROBLEM STATEMENT

In particular, the problem statement is specified below:

- Computational complexity replaces state complexity in the control plane.
- Data intelligence enables user choices and rewards innovations.
- Correlations between data analytics help to solve hard optimization problems.

## IV. PROPOSED SYSTEM

### 1. Data Driven Internet Architecture Design

Data Driven Design Network design shares much synergy with social and economical functioning in how the constituents interact and influence each other. From the perspective of economics, in a perfect competition market, consumers are more interested to purchase less cost and better products, while service providers are encouraged to improve productivity and supply better products. It is the transparent to economical data such as pricing and quality of products that enable vendor selection and resource optimization. The envision four essential planes of future Internet that enable the choice of network design alternatives and eventually lead to cost-effective, reliable, and agile network architecture.

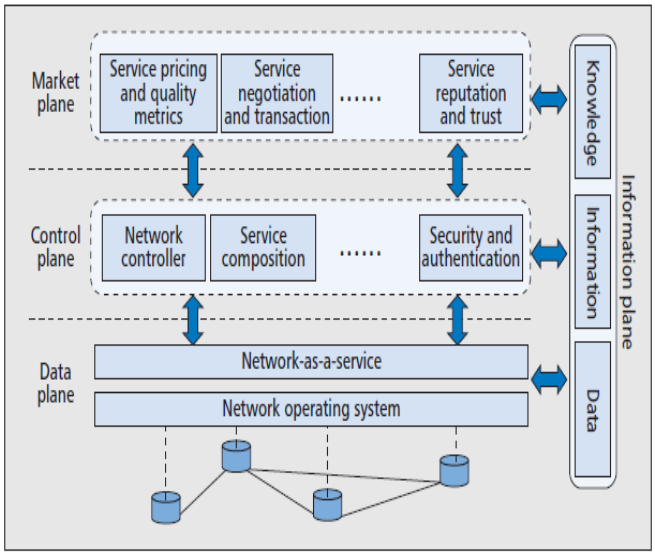


Fig 1: Data Driven Internet Architecture

As illustrated in Fig. 1, the four planes are control, data, information, and economics. The data plane sends packets and schedules flows based on rules, and the network operating systems and abstractions of network devices provides interface to the control plane.

The control plane runs centralized network control services to provide services such as forwarding, scheduling, and security policies are implemented onto the data plane. The information plane collects network measurement data and sends to the control plane, which uses real-time network snapshots for responding to significant events and statistical data for periodical tuning of network. The market plane hosts a marketplace where users, services, and infrastructures interact, allowing users to compare cost and quality, negotiate transactions, and pursue the most suitable matches to maximize the tangible metrics. The market plane requires information completeness, symmetry, and transparency is maintained in information plane. The knowledge learned from the information plane drives economical functions in the market plane is used in Internet architecture advancement.

The design of the future Internet will inevitably benefit to user based on cost, because choices can drive the competition and innovation necessary for future networks. The market plane is imposed on the network operators and service providers, requiring their investment in innovation and services. Users of the network can select from a range of alternative services that may differ in functionality, performance and cost. The choice mechanism also promotes the robustness and security of a network by

encouraging different solutions at all levels of the architecture. ChoiceNet is a recent example of how network architecture can provides core principle.

## 2. New Communication Models

Increasing computational power, outpacing network I/O speed, makes it possible to transform conventional communication models. Network operators progressively integrate computing power to alleviate the bottleneck of bandwidth resource in a network system. For example, in content distribution, application acceleration and load balancing can generate more revenue and optimize resource allocation of scarce network bandwidth. Using computing resources to compensate for network resources makes a communication model is more scalable and efficient on existing infrastructure. The Internet communication model has begun shifting from routing-centric to content/service/X-centric. Conventional routing centric communication models assume a "dumb data pipe" due to the scarcity of CPU and memory resources in network devices. With the increasing ubiquitous computing power present in the Internet, network controllers and middleboxes are exploiting such computational capabilities.

As shown in Fig.2, information centric networking (ICN) has changed how clients communicate with content providers. In ICN, content identifiers, rather than server IP addresses, become the handles of requests and replies.

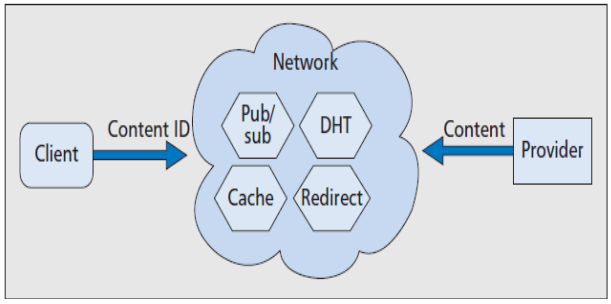


Fig 2: New Communication model

In such a way, ICN naturally supports various functionalities including content distribution, multicast, and mobility. DMap manages dynamic identifier- to-locator mapping for supporting mobility and content delivery.

## 3. Correlation Replaces Causality for Resource Management

Internet resource management problems are typically described with mathematical models, many of which are NP hard. The network topology of the Internet is too complicated to be measured and described. As a result, network management, task scheduling, and resource optimization are difficult to model with an ideal framework without assumptions far from reality. Even with such models, one often faces state explosion when attempting to solve them due to the distinct targets from multiple dimensions such as user requirements, network bandwidth, and geolocation. The stakeholders have objectives that may be competitive and adverse to each other, and each of these parties vies for its particular interest. Thus, it is difficult to discover the internal causality among all the competing factors in the Internet. It is desirable to consolidate the objectives and obtain an effective solution as opposed to seeking the optimal solution in an unmanageable time span, because benefits from new resource management strategies are immediate and bring new subsequent changes in user behavior, which may invalidate the original problem formulation. To describe the implications of such transformation using a server placement example as follows. An effective media server placement strategy is a classical facility location problem with the aim of choosing *M* replicas or hosting services among *N* potential sites, as shown in the upper part of

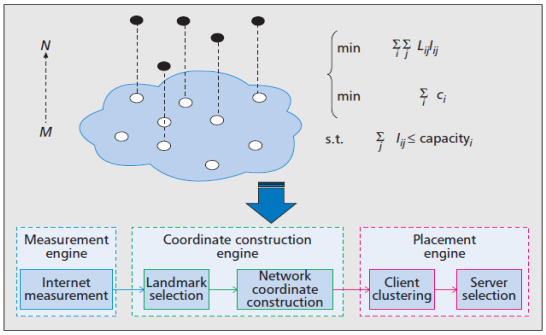


Fig 3: Change in scientific paradigm

Fig.3. It is typically modeled as a multi-objective optimization problem to solve, with the objectives of minimizing total costs and delay. Such models suffer from three major limitations:

• The assumption of a fixed candidate pool is invalidated by the rapid growth of cloud computing platforms and data centers, which introduces choices of pervasive hosting resources.

• Most of the existing solutions do not scale well with the number of server candidates and the number of clients: the *K*-mean problem or a facility location problem is known to be NP-hard.

• Traditional solutions fail to discover inherent rules and potential trends, and thus cannot give insight on the choice of an ideal number of servers. Enabled by the correlation discovered through big data analytics, the reconsideration of the work-server placement example and propose a novel scheme called NetClust, shown in the lower part in Fig. 3. NetClust takes advantage of the latest network coordinate techniques to obtain global network information for server placement, and leverages a clustering algorithm to determine the correlation between deployment cost and service performance.

## V. METHODOLOGY

The Changes of Internet Architecture Computational Complexity Replacing State Complexity. The control plane is overwhelmed with network state information. For example, multiple routing processes on a router depend on huge control knobs: route metrics, access control lists (ACLs), policies, and so on. The network state information such as dynamic states in forwarding information bases (FIBs), port settings, policies, packet filters, and timers, and mutual dependencies, may cause detrimental effects: faults, instability, inconsistency, and so on.

Disruptive changes in the control plane determine how the future Internet can be designed in network management and operation. The ongoing transformation is evident in data plane abstraction, control plane abstraction, and computational network control.

SDN architecture decouples the control and data planes by abstracting network devices with flows and actions applied on flows. Such data plane abstraction construct uniform control protocols (e.g., OpenFlow) and network operating systems on top of which the control plane can be operated programmatically as opposed to via manual configurations. network intelligence and state are logically centralized in a software-based SDN controller, which maintains a global view of the network and can programmatically fulfill complex control functions. Increasing computational capability at the network control plane changes in how networks are controlled. Current distributed control planes challenges to change new sophisticated control functions because distributed algorithms and protocols makes difficult to state dynamics, resulting in poor extensibility. Incremental additions of ad hoc control components prevent on the right protocol. On the contrary, a centralized optimization algorithm is more robust, and configuration-free protocols does better reachability and allow fine-grained flow control. As a result, designing a centralized control plane is more desirable for implementing sophisticated control logics. This control plane is logically central and can be implemented with multiple instances on multiple hardware platforms to ensure its reliability.

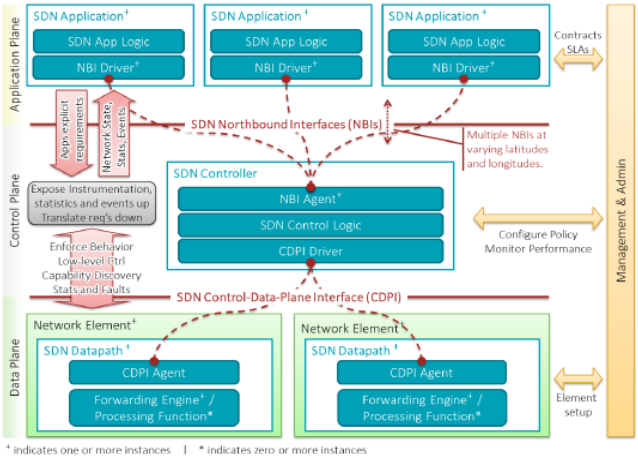


FIG 3: SOFTWARE DEFINED NETWORK

## VI. CHALLENGES FACED BY DATA-DRIVEN DESIGN PHILOSOPHY

### Data-Driven Architecture Design

The separation of the control and data planes has gained much traction. Most of the prior work focused on the performance of the data plane and enhanced features of the control plane until the emergence of SDN, which introduces horizontal partition of the packet processing procedures.

The success of SDN brings an interesting set of new problems such as network control verification and debugging, therefore calling for broader employment of computational intelligence in network architecture thanks to the advent of massive network measurement and monitoring data. It is challenging to derive the right abstraction of control plane and data plane, and leverage them to control network operations and states. At the information plane, the increasing volume and variety of network data and the knowledge derived from them (including network traffic variation, user access patterns, data center capacity and utilization, content generation and customization, etc.) reflect the natural interactions among user, content, and network. Many studies have shown that correlations learned from data can be utilized to effectively improve network resource allocation, reduce cost, and maximize revenue without resorting to formulation of NP-hard problems. Data-driven knowledge discovery of network operation and optimization guides the design of

future Internet. It remains an open question as to how to expose information in a transparent manner while preserving privacy. Data heterogeneity in time scale and units requires comprehensive data curation. Accuracy-tolerant data processing also needs effective approximation and learning.

The market plane serves as the arena where users reward services and technologies that are deemed innovative and helpful. From advertisement to content delivery, the Internet has been closely engaged with economical behaviors. However, economics-driven design is still in its infancy. The current architecture lacks transparent pricing, as well as transaction and selection systems for users and providers to match their interest in experiences, quality, and costs. In particular, the network core is far more closed than the edge with respect to service selection. The obstacles include the unwillingness of service providers to expose information, and users' inaccessibility to large selections of services with well defined marketable metrics. the open, transparent, fair marketplace for network technologies and services is established.

**1. Data Acquisition**

Data acquisition for information plane is required to address the volume, velocity, variety, and validation of big data. The objectives are to obtain sufficient data for understanding network infrastructure topology, network service performance, and user experiences, which are three crucial components. Progress has been made in data models and tools for at-scale network measurements. A network coordinate system models as a geometric space and characterizes the position of any node in the Internet by a coordinate in this space. IPlane utilizes the routing topology to build estimates the properties of the paths between arbitrary end hosts based on the path composition technique, which derives these estimates by composing the inferred properties of either path segments or links in the atlas. However, their data acquisition schemes are not ready for comprehensive, constant, or diverse measurements.

A big-data-based measurement platform named SITEWARE. By means of distributed network measuring; this platform orchestrates a variety of measurement tools, methods, and theories to evaluate network performance. Specifically, there is an awareness of network infrastructure with topology information (AS relationship, statistical features, etc.), network performance with quality of service (QoS) data (packet loss, latency, bandwidth, etc.), and user experience with QoE information (e.g., response time and service stability). From bottom up, our corresponding goals are to optimize the Internet architecture, protocol, and resource management, and task scheduling for improved user experiences. By using this layered measuring platform, a cross-layer method to systematically identify and address a fault. For example, if QoE is below expectation, there must be a problem in the corresponding QoS metrics. If a QoS metric is lower than a threshold, it implies that there must be an infrastructure-related problem, for example, changes in AS relationship or a bottleneck link between the source and destination.

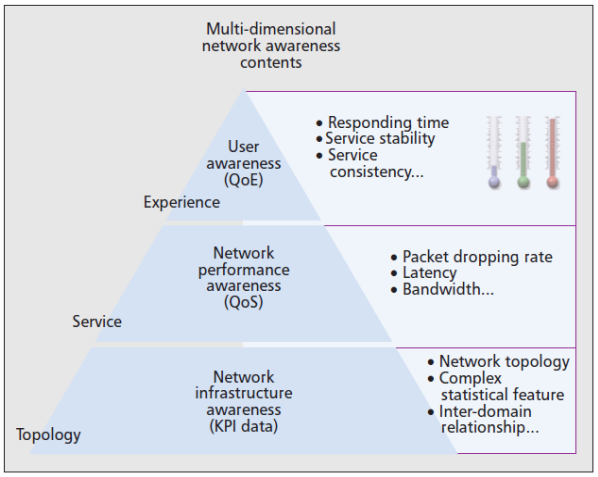


Fig 5: SITEWARE Platform for multi dimensional network measurement.

Technically, it is challenging to obtain complete, multi-dimensional, and accurate data due to its heterogeneity, scale, and dynamics in operations. Network malfunctions, failures, state changes, and mobility result in instability of data. Accurate and large-scale data collection remains a major undertaking because of the dynamics of operations. Addressing issues such as data validation, sampling rate, representativeness, and security is indispensable to Internet-scale data acquisition.

**2. Data Utilization**

As data obtained from Internet measurements feature multi-source, heterogeneity, inconsistency of entity, incompleteness, and inaccuracy, data curation needs theories and tools to support data representation, storage, integration, fusion, retrieval, and extraction.

Data volume keeps expanding, and faults and malfunctions occur frequently, complicated by the lack of a unified network information collection mechanism. The state information obtained by different operators or entities

needs to be shared and assembled for a global view of the Internet; however, their proprietary interfaces prevent the community from sharing data. Timely or even real-time responses to interactive inquiries about urgent Internet events are critical. A common ontology is desirable for mapping diverse data sources to measured objects and describing measurement data with accompanying metadata the structure of which can be defined with web ontology languages. Structured, semi-structured, and unstructured data about the Internet call for novel solutions to database indexing problems.

**3. Visualization:**

Data might be obscure and hide their potential value until they are made amenable to human perception. Data visualization exhibits the correlations and implications of raw data with images and dimensions so that our eyes can see and our brains can understand the trends and connections for us to act on. Example: 97 percent of the observable ASs in the Chinese Internet and their statistical features.

Nodes representing ASs in China are laid out across a series of two-dimensional concentric circles with their diameters inversely proportional to the corresponding coreness values, which imply the robustness of an AS. The sizes of nodes are exponentially proportional to the corresponding node degrees, and their colors help differentiate nodes of diverse coreness value.

**4. Predictive analytics:**

A data-intensive computing paradigm provides an effective way to identify the mingled factors in the Internet by searching for correlations. Utilizing various machine learning and data mining algorithms, one can derive knowledge from the data to support decisions on resource management, task scheduling, production recommendation, anomaly detection.

**5. Actionable instructions:**

The value of data depends highly on its usefulness in producing actionable items. These actions are directly tied to the benefits to users, ISPs, content providers, and all of society. They are intended to improve Internet performance metrics, management and control mechanisms, resource allocation policies, and so on. The complex and sometimes counter-intuitive relationships discovered from predicative analytics should furnish tangible mechanisms for Internet operators to work with, just like the interest rate and currency supply in macroeconomics.

## VII. CONCLUSION

Big data brings unprecedented opportunities for reshaping the future Internet. A data-driven approach takes advantage of the massive data and ubiquitous computing capabilities to transform network architectures, communication models, and resource management. Centralized network control such as in SDN replaces distributed and autonomous subsystems. Emerging communication models such as content-centric networks compensate for inadequate network bandwidth with computational intelligence. Discovery of the correlations among the Internet components allows for scalable resource optimization, an inherently hard problem. Data-driven network design calls for the effort to address data acquisition, curation, and utilization, and, most important, the Internet economics where transparent data empower choices of network technologies and services. There are abundant open issues in theories, mechanisms, and tools for harnessing data and fostering knowledge in this big data era, the ultimate goal of which is to derive actionable items for network design in the coming decades.

## REFERENCES

[1] M. Casado et al., “Fabric: A Retrospective on Evolving SDN,” HotSDN ’12, Aug. 2012, pp.85–90.

[2] ONF, “Software-Defined Networking: The New Norm for Networks,” Apr. 2012.

[3] D. S. Han et al., “XIA: Efficient Support for Evolvable Internetworking,” NSDI ’12, Apr. 2012.

[4] X. Liu et al., “A Case for A Coordinated Internet Video Control Plane,” ACM SIGCOMM ’12,Aug.2012,pp.359-70.

[5] A. Balachandran et al., “A Quest for an Internet Video Quality-of-Experi¬ence Metric,” HotNet’12,Oct.2012, pp. 97–102.

[6] T. Wolf et al., “Choice as a Principle in Network Architecture,” ACM SIG-COMM ’12, Aug. 2012, pp. 105–06.

[7] D. S. Han, Supporting Long Term Evolution in an Internet Architecture, Carnegie Mellon Univ., 2012.

[8] B. Ahlgren et al., "A Survey of Information-Centric Networking," IEEE Commun. Mag., vol. 50, no. 7, July 2012, pp. 26–36.

[9] T. Vu et al., "DMap: A Shared Hosting Scheme for Dynamic Identifier to Locator Mappings in the Global Internet," IEEE ICDCS'12, June 2012, pp. 698–707.

[10] H. Yin et al., "NetClust: A Framework for Scalable and Pareto-Optimal Media Server Placement," IEEE Trans. Multimedia, vol. 15, no. 8, Dec. 2013, pp. 2114–24.

[11] T. S. E. Ng and H. Zhuang, "Global Network Positioning: A New Approach to Network Distance Prediction," IEEE INFOCOM '02, June 2002, pp. 170–79.