MobilityFirst: A Clean Slate Network Architecture for Next-Generation ("5G") Mobility Services[1]

MobilityFirst Future Internet Architecture (FIA) Project Team
WINLAB, Rutgers University,
Contact: D. Raychaudhuri, ray@winlab.rutgers.edu

Abstract— This white paper presents an overview of wireless access considerations behind the design of the clean-slate MobilityFirst next-generation ("5G") mobile network architecture being developed under the NSF Future Internet Architecture (FIA) program. The MobilityFirst architecture is motivated by a historic shift of the Internet from the fixed host-server model to one in which access from mobile platforms becomes the norm. This implies the need for a future Internet protocol stack designed to handle the special needs of mobility services efficiently and at large scale. A number of key wireless access network requirements, including user/network mobility, varying wireless link quality and disconnection, multi-homing, ad hoc networking, flexible autonomous system boundaries, and spectrum coordination are identified along with a brief discussion of their implications for protocol design. This is followed by a summary of the MobilityFirst protocol design based on separation of names and locators, global name resolution service, storage-aware routing with hop-by-hop transport, integrated spectrum management, along with an edge-aware inter-domain routing framework. Illustrative examples showing how the MobilityFirst protocol stack supports mobility, multi-homing, and inter-network spectrum coordination are also given. The discussion concludes with a brief view of ongoing prototyping and validation efforts including experimental deployments of the MobilityFirst protocol on the national-scale GENI (Global Environment for Network Innovation) testbed.

1. INTRODUCTION

In 2010, the US National Science Foundation (NSF) initiated the Future Internet Architecture (FIA) program [1] which aims for a clean-slate redesign of the Internet to address emerging service and security needs for 2020 and beyond. Four team projects with distinct visions and technical approaches were supported under FIA – these include NDN, NEBULA, XIA and MobilityFirst (see [1] for links to individual project sites). The MobilityFirst architecture introduced here is based on the belief that the Internet is fast approaching an inflexion point with

---

[1] Research supported by NSF Future Internet Architecture (FIA) grant CNS-1040735

wireless/mobile devices overtaking wired PCs as the primary end-user device, i.e. mobility as the norm. Since the iPhone was introduced in 2007, worldwide smartphone usage continues to grow at an exponential rate – the Cisco VNI Global Mobile Data Traffic Forecast 2013 [2] predicts that traffic from smartphones alone will account for about 7.5 Exabytes/month in 2017, a factor of ~10x relative to 2013. The Cisco report also forecasts that "by 2016, wired devices will account for only 39% of all IP traffic". This fundamental shift in Internet usage represents a unique and timely opportunity to design a clean-slate architecture that takes into account the special requirements of mobile devices and applications. The goal is to develop a next-generation IP protocol stack which supports the anticipated scale of tens of billions of mobile devices as "first-class" network entities with excellent end-user/application performance, network efficiency and trustworthiness.

Existing protocol solutions for cellular mobile data service involve two sets of protocols (3GPP and IP) and the use of protocol tunnels and gateways that constrain the network topology and result in processing bottlenecks. Further, the current TCP/IP protocol framework has several limitations when applied to wireless access scenarios with mobile endpoints. IP address assignment and management via protocols such as DHCP and DNS are relatively static while TCP assumes the existence of a contemporaneous end-to-end path. In addition, IP addresses serve the dual roles of end-point identifier and routable network locator, making it difficult to deal with many aspects of dynamic mobility such as disconnection or multi-homing. Mobile IP provides an incremental solution to the dynamic address assignment problem through indirection to a home agent, but has limitations which have discouraged its widespread adoption. Proposals such as the Host Identify Protocol (HIP) [3] tackle the issue of separating names from addresses, but fall short of providing a complete solution for mobility scenarios, for example, heterogeneous cellular networks and machine-to-machine (M2M) communications. A more holistic approach is required to serve the full range of mobility platforms and applications, considering factors such as disconnection tolerance, content and context support, authentication, and privacy. Recent industry activity on "5G" wireless systems offers an opportunity for a clean-slate redesign of the mobility protocol stack to be fully harmonized with next-generation Internet protocols while also meeting the needs of the full set of anticipated applications such as mobile data, IoT (Internet-of-Things), content delivery, vehicular and cloud services.

The MobilityFirst (MF) architecture [4] is aimed at the above requirements using a set of protocol components very different from today's TCP/IP. Among the important design decisions are to separate names of network-attached objects from routable network addresses (locators), and to provide a fast global mechanism for dynamic binding between names and their current address(es). A second key feature of the design is the use of public key names (known as globally unique identifiers or GUIDs) to secure user, device, and network names without a single root of trust. Network services constituting the “narrow waist” of the protocol are defined in terms of GUIDs, making it possible to design a variety of advanced services such as dynamic mobility, disconnection tolerance, multicast, anycast, content retrieval, and context-based message delivery. The proposed network also uses robust storage-aware routing techniques along with hop-by-hop packet transport to provide a seamless solution across a range of wired and wireless access scenarios. The architecture introduces a separate management plane which enables decentralized visibility of network resources and supports more general forms of service level agreements between network entities. Finally, an optional computing layer at the routers is introduced to enable service customization and security/privacy processing capabilities inside the network.

In this paper, we focus on the application of the MobilityFirst protocol to wireless access networks. Sec. 2 briefly presents a view of how cellular-Internet convergence is enabled by the proposed mobility-centric protocol architecture. This is followed by a discussion of the wireless access challenges and service requirements (Sec. 3) that guided the design of the MF protocol stack. The MF protocol design is then outlined in Sec. 4, and major protocol components such as the global name resolution service (GNRS), storage-aware and delay tolerant routing, edge-aware inter-domain routing and content/context services are described. An overview of ongoing MobilityFirst prototyping and validation efforts is given in Sec. 5, including deployment on GENI. Concluding remarks are given in Sec. 6.

2. CELLULAR-INTERNET CONVERGENCE

As Internet-connected mobile devices will soon outnumber fixed PCs, a convergence of business models and technical standards associated with cellular networks and the Internet may be expected over the next decade. This process has already started, with cellular standards embracing the concept of “flat” IP-based networks without centralized gateways. In 4G/LTE, the cellular access network architecture has been significantly flattened with only a single specialized MME (mobility management entity) in the control path and SGW (service gateway) in the data path, and with commodity routers everywhere else in the network. The 3GPP architecture uses IETF protocols such as Mobile IP for wide-area mobility management, supplemented by micro-mobility, authentication and other services provided by the cellular MME. The future “5G” network specification is expected to move further in the direction of harmonization with the Internet Protocol (IP), and is thus a potential candidate for clean-slate redesign along the lines proposed here.

In our view, the next logical step in this direction is a completely flat mobile network architecture with native support for basic services such as authentication, dynamic association and handover, inter-network roaming, and disconnection tolerance. As shown in Fig.1, all routers, base stations and access points in the network would run the same future IP protocol stack, and clients would have a single identifier and uniform service APIs, making it possible to simply “plug in” wireless access technologies such as GSM, 3G/CDMA, LTE, WiMAX or Wi-Fi without requiring gateways. Such a uniform “mobile Internet” protocol solution across wired and wireless network technologies will eventually lead to convergence of cellular and Internet standards as both industries are serving the same mobile end-users. Beyond mobile data, any new protocol architecture should also support the requirements of emerging machine-to-machine (M2M) communications between embedded sensors, vehicular networks, and Internet-of-Things devices, which are expected to grow significantly over the next decade to an estimated 1.5 billion devices by 2017 [2].

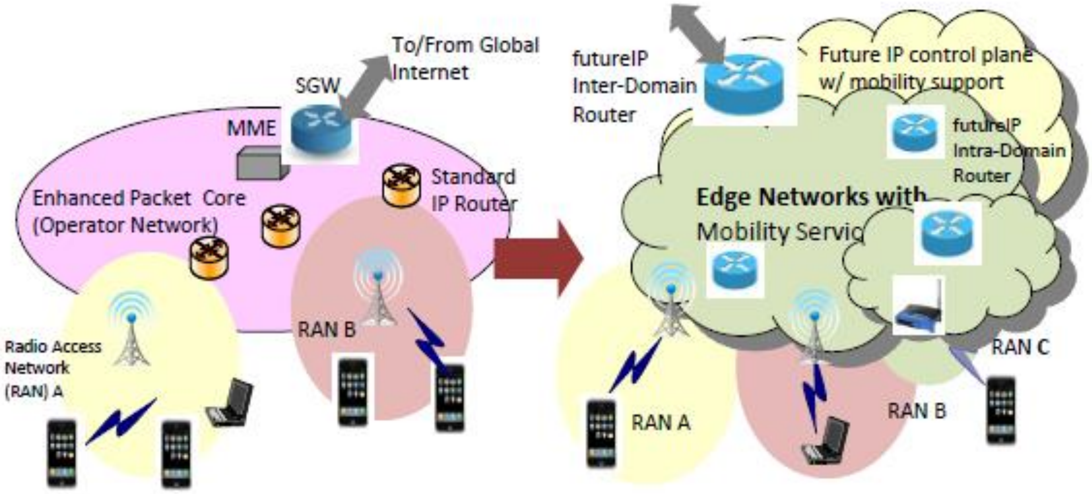


Fig. 1: Cellular-Internet Convergence Scenario

We note that a unified mobile Internet architecture is useful to both cellular network operators seeking to improve performance, as well as to more general Internet service providers (ISPs) aiming to introduce mobility services across heterogeneous access networks. For example, an ISP that currently offers standard Internet access service could expand the offering to include seamless mobility across multiple wireless networks such as Wi-Fi hot-spots using standard network element (router, base station, access point) capabilities without the need for a specialized control framework. This type of heterogeneous wireless access service is sometimes referred to as “open wireless networks” [5] in which loosely coupled access networks use a common protocol to support basic mobility needs such as authentication, handover and inter-network roaming. Such access networks may be expected to become a viable alternative to managed cellular services if they are able to offer a level of access and mobility that could be adequate for some portion of end-users and applications. Cellular providers incorporating Wi-Fi hot-spots and 3G/4G/5G small cells to supplement their existing macro-cellular deployments could also use the same flat future IP protocol to provide mobility services across these heterogeneous networks

3. WIRELESS ACCESS CHALLENGES & REQUIREMENTS

A. Host and Network Mobility

The foremost characteristic of untethered nodes is that their points of attachment to the Internet can change easily and rapidly. The need for supporting mobility arises when an individual node or a group of nodes, for example a bus/train/plane network, moves and reconnects to the Internet. There has been extensive work on enhancing the Internet protocol suite to support mobility, most notably with mobility anchors as in Mobile IP. These solutions are based on a set of implicit assumptions – that users have an immutable “home” network, are connected to a single network at a time, and transitions across networks are infrequent. Consequently, packets in the current architecture are sourced from, and destined for, the network attachment point of end-hosts, i.e. their IP addresses. However, this network model has changed since Mobile IP was conceived. It is important to understand the simple but fundamental requirement for mobility support – hosts need permanent names irrespective of their attachment points, and the network needs a packet transmission primitive that employs permanent names. This functional requirement can be translated to the following protocol design requirements:

A1. Disambiguation of the dual-roles of an IP address as both an identifier and a locator into two different primitives - a permanent name and a network-specific temporary locator.
A2. Dynamic binding of names to network addresses/locators.

B. Varying wireless link quality

Fluctuations in access link quality are an intrinsic property of the wireless medium – achievable bit rates in both Wi-Fi and 4G systems, can show large variations within a fraction of a second and disconnection due to mobility and/or insufficient signal strength is not uncommon. While these variations are usually handled at the PHY and MAC layers, they invalidate some implicit assumptions in the control algorithms used in the Internet. For example, it has been long known that TCP congestion control treats wireless link errors as congestion losses and performs poorly in high variation wireless channels. Given the increasing dominance of the wireless last hop for Internet access, such link quality variations need to be natively supported at different layers of the Internet architecture. This leads to the following requirements:

B1. Link quality awareness at both the intra-domain and inter-domain routing layers to enable robust packet delivery strategies.
B2. Disconnection-tolerant routing and transport protocols that are capable of temporarily storing packets during disconnections and rerouting in-transit packets to new points of attachments.

C. Accessing multiple networks

A typical wireless device in an urban area today might see 3-5 cellular networks and 10-20 Wi-Fi access points, but accesses only one of these due to both technical and business model constraints. Current techniques supporting simultaneous use of multiple interfaces rely on enhancements to the underlying end-to-end transport layer (see [6] and references therein). Specifically, these mechanisms require a multi-homed end-point to inform the sender about its multiple interfaces prior to the commencement of data-flow, and a data-striping algorithm on the sender stack that adapts the packet rate of each interface. This results in rigidity in two key aspects: (i) There is no mechanism by which users can specify under what conditions, and in what manner the interfaces are to be used; (ii) Since all decision logic is implemented only at the end-nodes, in-network routers cannot adapt or buffer the flows in accordance with wireless channel quality variations. Thus efficient support for host multi-homing induces the following key requirements:

C1. Support for binding a single name to multiple addresses and interfaces.
C2. A routing plane capable of modifying the data-striping and storing decisions in accordance with the link quality at each interface.
C3. Service semantics to support interface selection and utilization (e.g. “send to all interfaces”, “send to higher-throughput interface”, “send only to Wi-Fi”, etc.).

D. Ad hoc networks

Wireless ad hoc networks are important for infrastructure-less vehicle-to-vehicle (V2V) and sensor network scenarios, last-mile connectivity and applications such as photo/video sharing, local social networking, and multi-player gaming. One view of Internet design is that ad hoc networks are just a type of edge network; as long as they are connected to the Internet via a boundary IP router, the protocols used within the ad hoc network can be ignored. However, the ubiquity of non-specialized devices requiring support for ad hoc networking (e.g. phones, tablets, laptops, vehicular infotainment systems, etc.) forms a strong argument for an integrated design that avoids boundary translation solution. Integration of such networks within the framework of a future Internet design results in the following distinct requirements:

D1. Critical network services such as authentication and dynamic binding of names to addresses should be capable of disconnected-mode operation.

D2. Routing and transport protocols should be robust to opportunistic association and changing network topologies.

E. Content and context addressability

Along with the shift from fixed to mobile nodes, the Internet is becoming content and context-driven. Content-driven usage refers to the retrieval of specific content (as opposed to communicating with a specific destination), while context-services use external conditions, including time, location, and network attachment, to deliver information to/from end-hosts. In these use-cases, it is necessary to use the content or context as a first-class primitive in packet transmission, i.e. it should be as easy to use content/context semantics such as “fetch content X from nearest source” or “send to all nodes at location Y,” as the traditional end-to-end semantic “send to address Z.” Supporting these use-cases in mobile scenarios leads to the following requirements:

E1. The architecture should enable dynamic identification of endpoints based on content/context attributes.
E2. Since the context attributes of mobile nodes can change rapidly, there is a requirement for fast mechanisms that capture the context and make it available as a packet delivery primitive.

F. Spectrum Access Coordination

Finally, a critical challenge that differentiates wireless networks from wired networks, but which is common across all forms of wireless networks – cellular LTE, Wi-Fi, white-space networks, etc. – is the need for devices to coordinate their use of spectrum. These coordination schemes, whether centralized, distributed, or a hybrid, are typically implemented through overlay channels – for example, the IETF PAWS protocol for accessing white space database uses an HTTPS overlay [7], and the X2 interface between LTE base stations uses SCTP over IP [8]. However supporting these wireless control plane functions at the scale of thousands of devices/km requires an integrated approach satisfying the following requirements:

F1. Support for a low-latency control plane that is unaffected by data plane congestion.
F2. Dynamic multicast of control messages, based on geographic location and radio-range of the sender, to enable efficient distributed coordination schemes.

Although we do not focus on security aspects in this paper, the requirements of location privacy, strong authentication of ownership, mechanisms against mobility spoofing attacks and fast authentication mechanisms must also be taken into account for a mobile-centric future Internet architecture.

4. MOBILITYFIRST PROTOCOL ARCHITECTURE

The MobilityFirst architecture [4] is based on the idea of separating “names” of end-users or other network-connected objects, and their routable addresses or locators. As shown on the top left of Fig. 2, the name-based service layer uses flat public keys as globally unique identifiers (GUIDs) for all network attached objects, whether a device, content file or context item. A GUID can be assigned to a network object by one of multiple name certification services (NCSs), and is derived through a cryptographic hash of the public key that corresponds to that object. The GUID being directly derived from the public key gives it a self-certifying property; authenticating a node does not require an external authority [3]. This is essential when communication to a third-party server is not possible or introduces excessive delay to critical applications. GUIDs assigned to network objects are mapped to a set of network addresses (NAs) or locators corresponding to current points of attachment. The dynamic mapping of GUIDs to NAs is made possible through a logically centralized, but physically distributed infrastructure called the global name resolution service (GNRS). The MF architecture also incorporates storage-aware intra-domain routing and edge-aware inter-domain routing. As shown in the figure, routers implement connectionless packet forwarding using both names and network addresses. The GUID name is considered to be the authoritative packet header, while routers may optionally use NAs for “fast path forwarding” after the GUID to NA bindings are determined at the first router along the path. In order to deal with dynamic connectivity changes, another option is late-binding in which a router may choose to rebind the GUID with NAs while a packet is in transit. Routers in MF also provide in-network caching of named content, and have an optional compute-layer to enable customized pluggable services such as media transcoding or privacy.

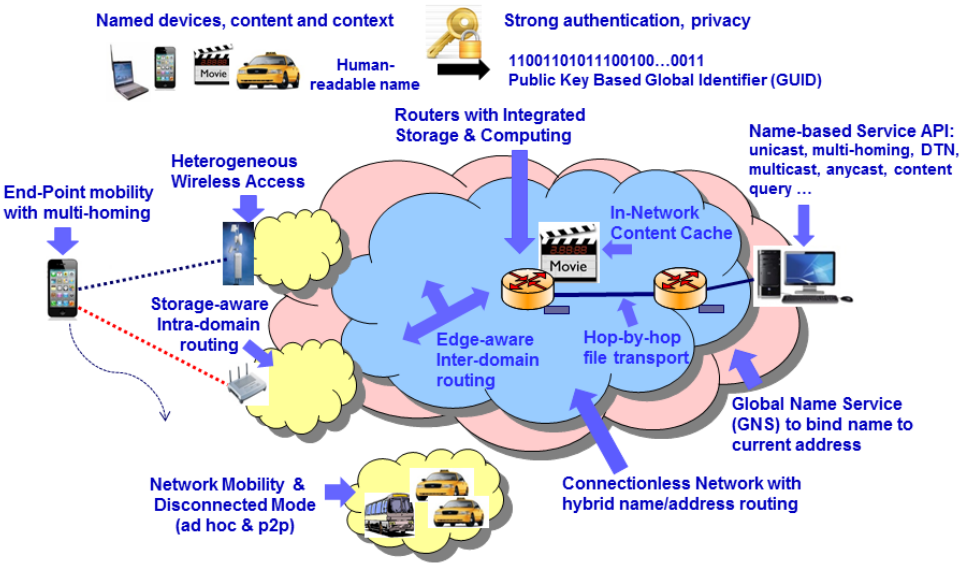


Fig. 2: Overview of MobilityFirst Architecture

In the following subsections, we present further details on key architectural components. Table 1 summarizes how these MF protocol components achieve the set of wireless access requirements identified in Sec. 3.

A. Dynamic Name-Address Bindings

When a user sends packets directed to a particular identifier (GUID), the network must quickly ascertain the set of locators (NAs) attached to the GUID and route the packets correspondingly. We address the challenge of providing a fast global name resolution service at Internet scale through a router DHT-based *Direct Mapping (DMap)* scheme for achieving a good balance between scalability, low update/query latency, and availability [9]. In order to perform the name resolution for a given GUID, DMap distributes the storage of GUID:NA mappings amongst Internet routers using an in-network single-hop hashing technique which derives the address of a storage router directly from the GUID. In a detailed simulation study [9], it was shown that DMap achieves a 95th percentile round trip query response time under 100ms.

Figure 3 shows how a packet is processed in the MobilityFirst network. The service API specifies the destination GUID along with service options such as anycast, multicast or multi-homing. The first router in the network obtains the current NA bindings corresponding to this GUID by accessing the global name resolution service and the returned NA values are then appended to the packet header. Subsequent routers in the network can then forward packets based on the specified NAs. In the example shown in Fig. 3, the network object ("John's devices”) is a multi-homed compound object with two devices represented by a single GUID. The dynamic mapping of GUID to NA provided by the GNRS thus enables various services including basic mobility across networks (the GNRS simply updates the current NA values), disconnection (routers hold the packet and query the GNRS periodically for a new value of NA), and multi-homing (the GUID returns multiple NAs and network routers forward and multicast the packet towards the specified networks).

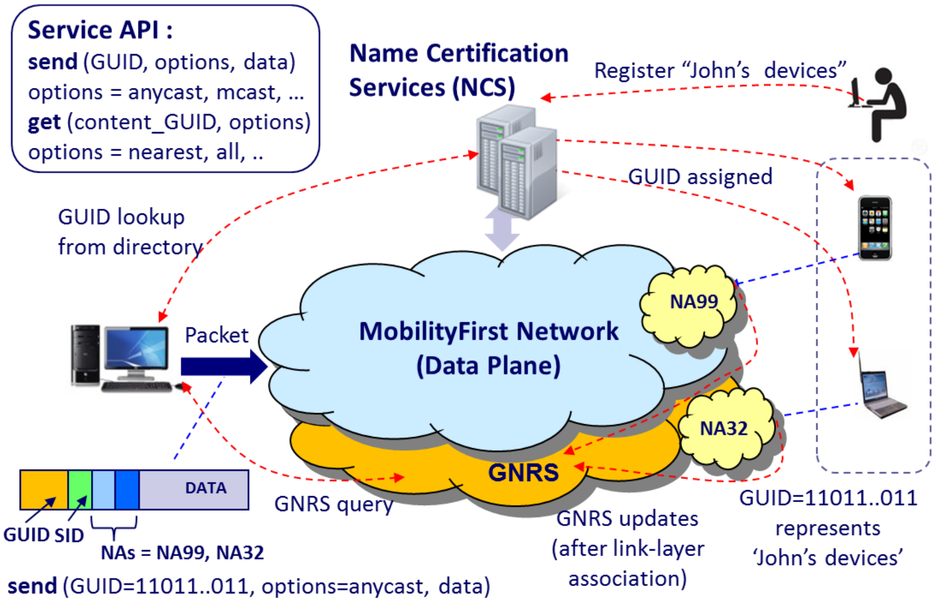


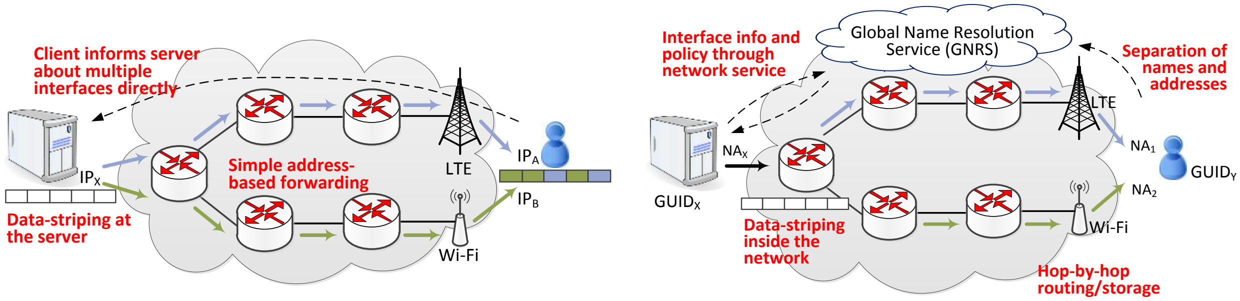
Fig. 3: Mapping of GUIDs to Network Addresses Using Global Name Resolution Services

While the dynamic mapping of GUIDs to NAs through a logically centralized in-network service is at the heart of MF’s mobility solution, the functionality of the GNRS goes beyond the basic name-to-address translation operations. In order to embed context-awareness and user policies (requirements E1 and C3) in the architecture, the GNRS maintains a list of attribute-value pairs for each GUID. Some examples are {geolocation: [lat, long], type: “host”, multi-homing policy: “send to all available interfaces”}, and each value can recursively encode lower-level attributes and lists of values, similar in spirit to JavaScript Object Notations. To support the formation and operation of ad hoc networks, MobilityFirst provides a boot-strapping mechanism through which nodes can invoke an ad-hoc GNRS (aGNRS) mode and directly share the list of GUIDs among the members of the ad hoc network [10].

B. Storage-aware and Delay Tolerant Intra-domain Routing

MobilityFirst uses a generalized storage-aware routing (GSTAR) algorithm to support delay and disruption tolerance in the routing layer. In GSTAR, each router employs in-network storage that facilitates store vs. forward decisions in response to varying link quality and disconnections [11]. These decisions are based on both short-term and long-term path quality metrics. In addition, packets along paths that become disconnected are handled by a disruption tolerant networking (DTN) mode of the protocol with delayed delivery and replication features. In particular, each router maintains two types of topology information: (i) An intra-partition graph is formed by collecting flooded link state advertisements which carry fine-grained, time-sensitive information about the intra-network links; (ii) A DTN graph is maintained via epidemically disseminated link-state advertisements which carry connection probabilities between all nodes in the network. Recent results [11] indicate that by intelligently utilizing in-network storage, GSTAR outperforms traditional and storage-augmented link-state protocols in both wired and wireless network environments.

In MobilityFirst, the requirements of multi-homing are met by incorporating support for multi-homed nodes directly in the routing layer (as opposed to the current end-to-end approach). As shown in Fig. 4, MF multi-homing makes use of network-assistance in two important aspects. First, the GNRS is used by multi-homed nodes to specify the availability of multiple interfaces and the corresponding interface preference policies. Second, the task of data-striping is shifted from the end-host stack to the in-network routers which have a better view of the end-to-end path quality through the underlying routing layer.



(a) End-to-end Multi-homing in IP based networks (b) Network-assisted Multi-homing in MobilityFirst

Fig. 4: Multi-homing approach used in MobilityFirst, and comparison with IP

C. Edge-aware Inter-domain Routing

While the wireless edge network properties are typically considered irrelevant to the design of the inter-domain routing protocols, the Edge-aware Inter-domain Routing (EIR) protocol used in MF was motivated by awareness that mobility at the edge has implications for routing between networks. Some of the resulting requirements cannot be met by current Border Gateway Protocol (BGP) solutions; see [12] for a detailed comparison between EIR and BGP. Since BGP does not differentiate between wired and wireless inter-network links, routing decisions based on capacity constraints are difficult. For example, in an early in-flight Wi-Fi implementation, Boeing associated each flight with an IP address block that was announced to the global routing system from different locations during the flight [13]. Networks receiving such announcements were unaware that the last hop for this path had a ground-to-plane wireless link instead of a high-capacity peering-point wired link and thus could send excess traffic towards this network. The key improvement of EIR is the propagation of aggregated link-level information in the inter-domain routing. In EIR, coarse-grained link-level information about each inter-network link is propagated through the routing protocol to enable forwarding decisions based on aggregate edge network properties. Additionally, each AS has the option of exposing the internal topology of its network through which other ASs can take fine-grained link-quality aware routing decisions. The ensuing increased routing update overhead is handled by adaptively damping the updates as they traverse through the network; a 26000 node trace-driven simulation shows that the resulting overhead is bounded and well below the observed global BGP overhead [12].

D. Content and Context Services

In MobilityFirst, content is a first-class endpoint principal that is represented using GUIDs in the same manner as interfaces or devices. Content providers create GUIDs for their content and insert an entry into the GNRS denoting its network address and the content GUID. A consumer retrieving this content first obtains its GUID through a well-known name assignment service and sends a *get('GUID')* primitive to the network along with its own network address. The first router queries the GNRS to resolve the GUID to a network address and relays the query to the provider. The content on its path from the provider to the consumer can optionally be cached at any router by its GUID. Future queries for the same GUID received by one of these routers triggers the router to send back a cached copy of the content and the packet is terminated without being sent to the original content provider. The key advantage of such an approach over the traditional endpoint-based IP communication is that when the same content is available at multiple destinations, the GNRS can respond to content-GUID queries with the address of the closest known source.

Context services are enabled through the combination of attribute-value pairs and GUID-to-GUID indirection in the GNRS. As mentioned, each device can register context information through a client-side agent that updates the GNRS entry tied to the device; for example, a mobile device might choose to update the *geolocation* attribute. To do so, the client sends an update message to the GNRS whenever the mobile moves more than a desired-threshold distance away from the previously updated position. Context-aware delivery ("send this message to all devices within a geographic location") is implemented by performing a search over attributes defined in the context-specification of the delivery-primitive (in this case over *geolocation*.) The GNRS then returns the set of NAs corresponding to the GUIDs that match the context specified by the user. In the mobile scenario, this method of dynamically defining a multicast group based on contextual information simplifies the implementation of location-based services and offers several new possibilities, including spectrum management as explained next.

E. Management Plane Services

The MF management plane facilitates spectrum coordination through dissemination of spectrum usage information to networks within radio interference range. In this architecture, routers directly connected to the base stations or access points use the GNRS to dynamically multicast control plane messages originating from a source device to the set of potential radio-interferers. As illustrated in Fig. 5, the source X of any spectrum management message, signs it (L_x, r_x) where L_x is the geolocation of X and r_x is the radius of operation obtained by equating: $PL_x(r) = P_{x,max} + G_x - S_{x,min} - N,$ where PL_x is the appropriate indoor/outdoor pathloss model used, $P_{x,max}$ is the maximum transmit power of X, $S_{x,min}$ is the minimum received power required for operation and N is the noise floor. Upon receiving this message, the router performs a GNRS query to find the address of all other registered devices operating in the same frequency band and spatial region. The resulting distributed spectrum management service allows for co-operation between independent access networks. For example, two virtual access networks using physically overlapping sets of Wi-Fi AP's can coordinate their spectrum usage in order to reduce interference. An example evaluation for overlapping Wi-Fi grids (as in stadiums or dense urban areas) shows 150-200% throughput improvements for clients most affected by the interference [14].

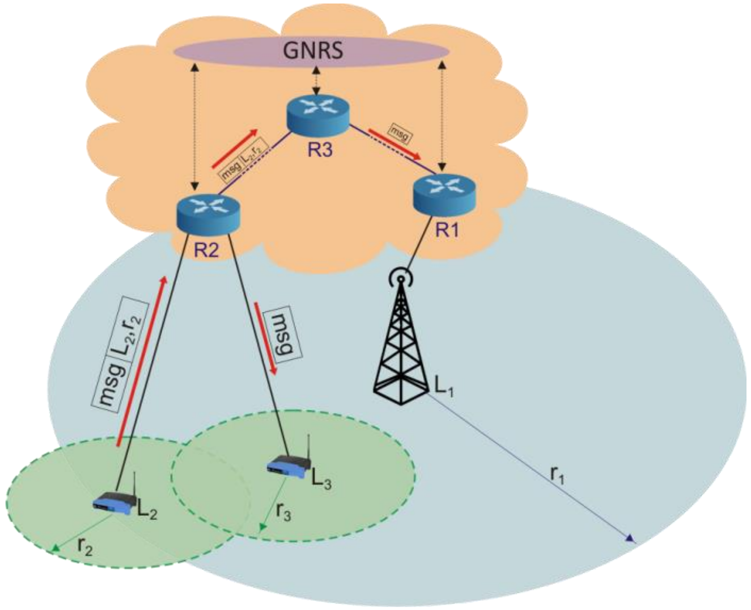


Fig. 5: Management plane service for spectrum coordination

Identified Wireless Access Requirements	MobilityFirst Protocol Elements				
	Global Name Resolution Service	Storage-aware Intra-domain Routing	Edge-aware Inter-domain Routing	Content & Context Services	Mgmt. Plane Services
A1. Identity/Location Separation	✓				
A2. Dynamic Binding	✓				
B1. Link Quality Awareness		✓	✓		✓
B2. Disconnection Tolerance		✓	✓		
C1. Multiple Addresses	✓				
C2. In-transit Decisions		✓		✓	
C3. Multi-homing policy semantics	✓				
D1. Disconnection-mode operation	✓				
D2. Topology Robustness		✓	✓		
E1. Content/Context Identification	✓			✓	
E2. Fast Updates				✓	
F1. Low-latency control plane					✓
F2. Dynamic distribution of control	✓				✓

Table 1: Key wireless access and mobility requirements and the MobilityFirst protocol elements that address each requirement.

5. PROTOTYPING & VALIDATION

The MobilityFirst Future Internet Architecture (FIA) project has employed the GENI testbed [15] for a series of tests and experiments over several years to assist in development and validation of the basic algorithms and protocols underlying the mobility-focused architecture. Key components of the MobilityFirst architecture selected for early validation in GENI include a massively scalable global name resolution service [9] and generalized storage aware routing [11]. This test configuration enables the evaluation of MobilityFirst protocols in a realistic setting that would be difficult to create without the use of GENI. The proof-of-concept network on GENI consists of 12 routers deployed on programmable GENI platforms (i.e. ProtoGENI nodes) spread across the US, with 3-4 edge networks (located at sites such as BBN, Cambridge, MA, WINLAB, Rutgers, North Brunswick, NJ and University of Utah, Salt Lake City, UT) having a WiMAX base station and/or WiFi access points for end-user mobile access. An additional node was deployed in Tokyo, Japan using the experimental FLARE programmable router platform recently developed by University of Tokyo researchers [16]. The corresponding topology (shown in Fig. 6) employs VLAN-based layer-2 stitching across multiple participating networks to establish a single layer-2 network across all deployed components. The configuration provides realistic wide-area RTT delays between router nodes, and combines a variety of link speeds and access technologies.

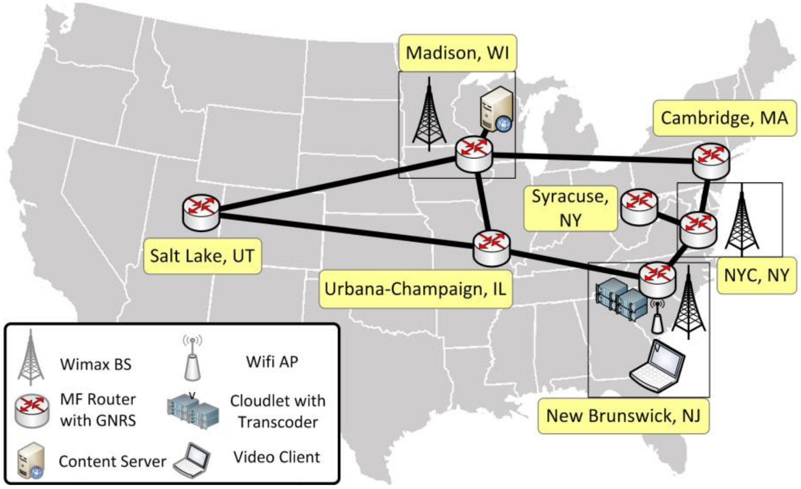


Fig. 6: Long-running deployment of MobilityFirst prototype network on GENI

The GENI setup outlined in Figure 6 has been used to successfully conduct a variety of validation experiments and proof-of-concept mobility service demonstrations including multi-homing (HotMobile 2012, GEC-17, 2013), content delivery with in-network media transcoding (GEC-20, 2014) and emergency response network with context-aware message delivery (US Ignite/GEC-22, 2015) to mention a few. The MobilityFirst deployment on GENI will also be used to support small-scale user trials for advanced mobility services in cooperation with an Internet Service Provider (ISP) in Madison, Wisconsin, planned for 2015-16.

6. CONCLUDING REMARKS

This white paper presented an overview of next-generation mobile Internet ("5G") design considerations driven by emerging wireless access and mobility scenarios. Key protocol requirements have been identified including name/address separation, robustness with respect to link quality variation and disconnection, multi-homing, ad hoc network formation, content/context addressability, and spectrum coordination. Design features of the MobilityFirst protocol stack have been outlined and shown to address these requirements. While comprehensive coverage of all design goals and protocol features is beyond the scope of this paper, we have aimed to provide a general understanding of the design approach and key features of the MF stack. The MobilityFirst protocol has been extensively validated using a combination of simulation, emulation and experimental trials – most recently, we demonstrated mobility service with a dual-homed smartphone over the GENI network with 12 MF routers across the US and multiple wireless access networks with 4G/Wi-Fi service. In the next phase of the project, we plan to conduct further laboratory and real-world trial evaluations of the proposed technology as applied to the cellular-Internet convergence scenario discussed here.

REFERENCES

[1] NSF Future Internet Architecture Project, http://www.nets-fia.net/.

[2] Cisco Visual Networking Index: Global mobile data traffic forecast update, 2012-2017, Cisco White Paper, Feb. 2013.

[3] R. Moskowitz and P. Nikander, Host Identity Protocol (HIP) Architecture, IETF Internet Standard, RFC 4423, May 2006.

[4] Dipankar Raychaudhuri, Kiran Nagaraja, Arun Venkataramani, "MobilityFirst: a robust and trustworthy mobility-centric architecture for the future Internet," *Mobile Computing and Communications Review,* 16(3): 2-13 (2012).

[5] R. D. Yates, and W. Lehr, "MobilityFirst, LTE and the evolution of mobile networks," in *Proceedings of* DYSPAN, 2012.

[6] J. Iyengar, P. Amer, and R. Stewart, "Concurrent multipath transfer using SCTP multihoming over independent end-to-end paths," *Networking, IEEE/ACM Transactions on*, vol. 14, no. 5, pp. 951–964, 2006.

[7] IETF Internet-Draft: Protocol to Access Spectrum Database, http://tools.ietf.org/pdf/draft-ietf-paws-protocol-05.pdf, June 2013.

[8] 3GPP TS 36.420, X2 general aspects and principles, http://www.3gpp.org

[9] T. Vu et al., "DMap: A Shared Hosting Scheme for Dynamic Identifier to Locator Mappings in the Global Internet," in *Proceedings of ICDCS,* 2012.

[10] A. Baid et al., "Enabling Vehicular Networking in the MobilityFirst Future Internet Architecture", in *Proceedings of IEEE WoWMoM,* 2013.

[11] N. Somani, A. Chanda, S. C. Nelson, and D. Raychaudhuri, "Storage- Aware Routing for Robust and Efficient Services in the Future Mobile Internet," in *Proceedings of ICC FutureNet V*, 2012.

[12] T. Vu, A. Baid, H. Nguyen, D. Raychaudhuri, "EIR: Edge-aware Inter-domain Routing Protocol for the Future Mobile Internet", *WINLAB Technical Report*, WINLAB-TR-414, 2013.

[13] B. Abarbanel, "Implementing global network mobility using BGP," NANOG Presentation, http://www.nanog.org/meetingarchives/nanog31/presentations/abarbanel.pdf, May 2004.

[14] A. Baid, M. Schapira, I. Seskar, J. Rexford, and D. Raychaudhuri, "Network cooperation for client-AP association optimization," in *Proceedings of RAWNET Workshop,* 2012.

[15] D. Raychaudhuri, "Using GENI to Prototype the MobilityFirst Future Internet Architecture", Mar 2013, http://www.geni.net/?p=2644.

[16] Aki Nakao, "Deeply Programmable Network Through Advanced Network Virtualization", Network Virtualization Symposium 2012, Keynote Talk, http://www.ieice.org/~nv/2nd-nakao.pdf