putting the cloud in the palm of your hand

Victor Bahl

5.23.2012

sad reality of mobile computing

hardware limitations

- vs. static elements of same era (deskt
- weight, power, size constraints
- CPU, memory, display, ke

finite energy

- actio erred
- costs energy

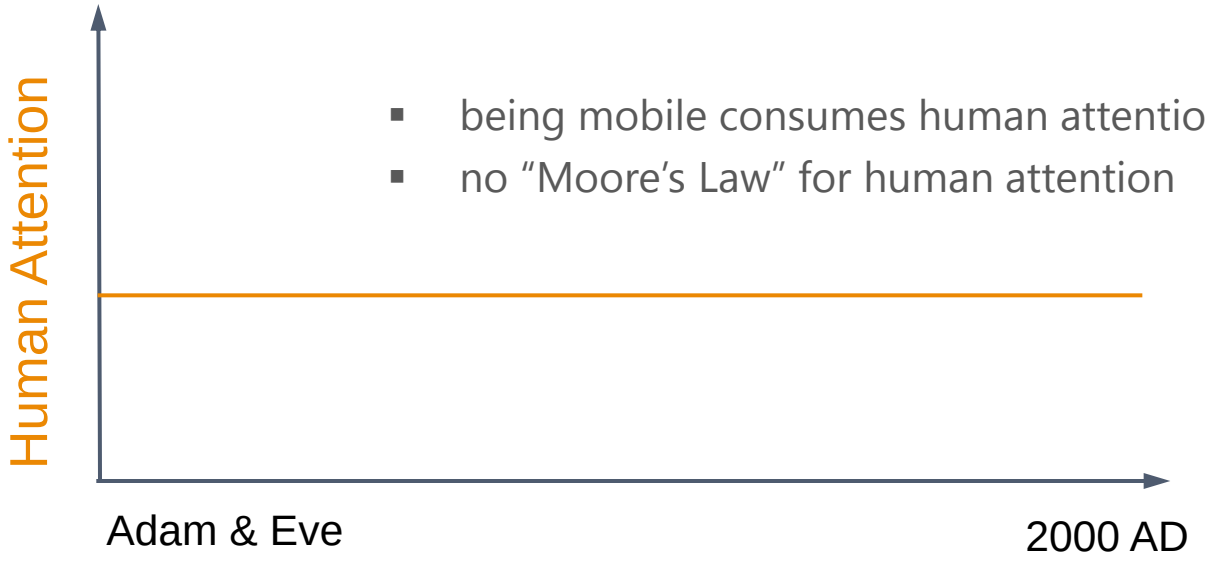
mmunication uncertainty

- ndwidth / latency variation
- intermittent connectivity
- may cost real money, require service agreements

True 15+ years ago (early 1990s)

- huge hardware & wireless networking improvements since but deep essentials still the same. Will the same slide will be true in 2020?

resource poverty hurts



- being mobile consumes human attention
- no "Moore's Law" for human attention

Clever exploitation needed to deliver benefits

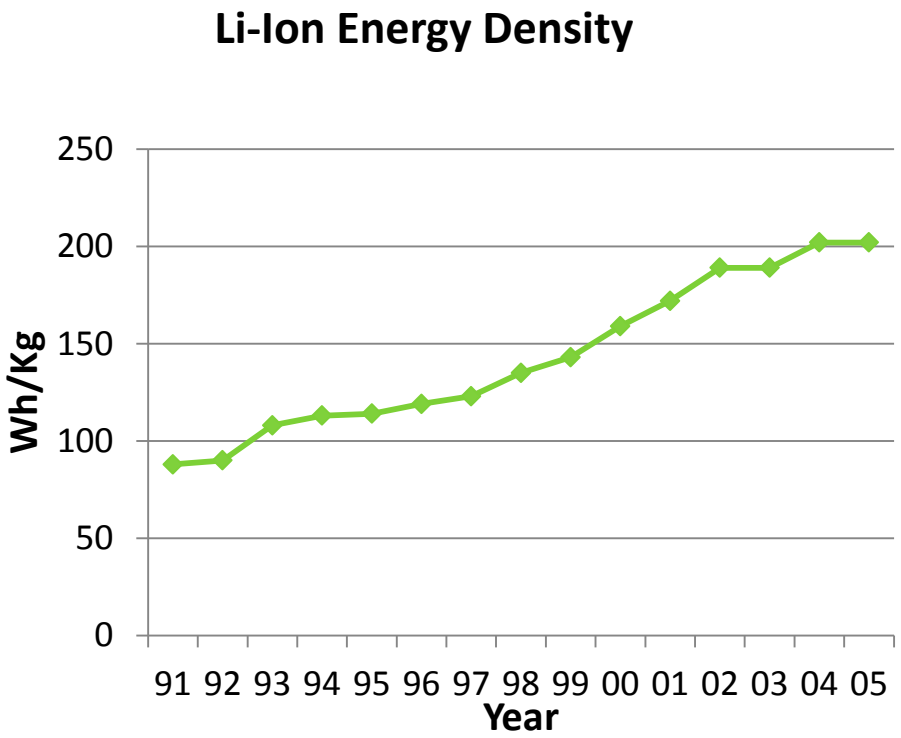
- machine learning, activity inferencing, context awareness
- natural language translation, speech recognition, …
- computer vision, context awareness, augmented reality
- reuse of familiar (non-mobile) software environments

…

→ Vastly superior mobile user experience

Courtesy. M. Satya, CMU

energy scarcity: silver bullet seems unlikely



lagged behind

- Higher voltage batteries (4.35 V vs. 4.2V) – 8% improvement
- Silicon anode adoption (vs. graphite) – 30% improvement

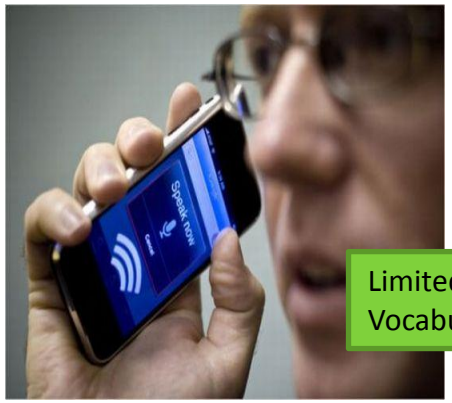
trade-offs

- Fast charging = lower capacity
- Slow charging = higher capacity

CPU performance improvement during same period: 246x

today's mobile apps are not reaching their full potential

Speech recognition & synthesis



Limited Vocabulary

Augmented Reality

Too CPU intensive

Healthcare sensing & analysis

Too Energy intensive

3D Interactive Gaming

Not on par with desktop counterparts

better together: phone + cloud

Phone offers ubiquitous connectivity and context awareness.

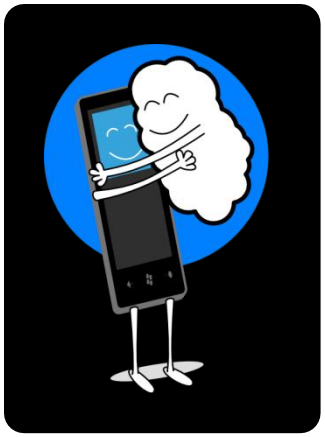
The cloud offers near-limitless resources

Azure

Together, they enable applications that were simply not possible before

vision: cloud in the palm of your hand

Enable mobile application developers to fully realize the potential of the cloud, and to do so quickly, reliably and easily.



from vision to strategy

getting to >100K cloud enabled apps

focus on the developer – provide programming support for

- → computational offload
- → resource intensive cloud services
- → multi-device programming

programming support for computational offload

Remote execution can reduce energy consumption and improve performance

opportunistic use of the cloud

research challenges

- what to offload?
- how to dynamically decide when to offload?
- how to minimize programmer effort?

important for adoption: a simple programming model

- app developer community has varying expertise & skills
  - Cannot require app developers to become experts in distributed systems

strategy

- developers build standalone apps with simple annotations but no changes to program logic
- use of nearby and cloud-server resources is opportunistic
- result: applications adapt as their execution environment changes

enabling simple program partitioning

Programming Model

- Dynamic partitioning made simple for th partitioning
  - Programmer builds app as standalone phone app
  - Programmer adds .NET attributes to indi "remoteable" methods / classes

```
[Remoteable]
ArrayList GetValidMoves(Square s)
{
    if (s.IsEmpty())
    {
        return new ArrayList();
    }
    if (s.Piece.IsEnemyOf(active))
    {
        //this piece does not belong to the active side, no moves possible
        return new ArrayList();
    }
    //forward the call to the Rule-class
    return rules.getMoves(s);
}
```

Salient Point:
The model supports disconnected operations

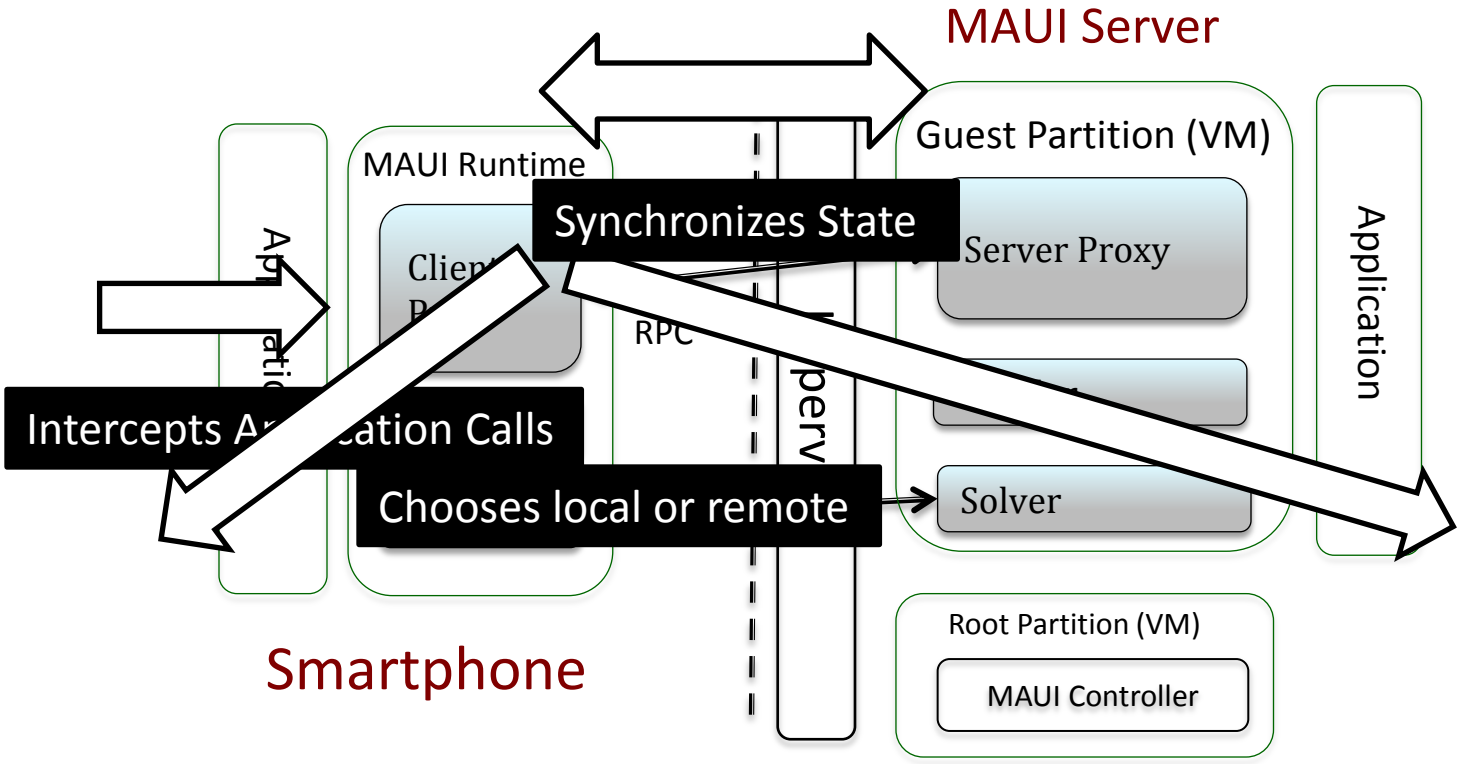
- Runtime: partitions (splits) the program at run-time
  - Can optimize for energy-savings, or performance

Why not use a static client/server split?

- Developers need to revisit application structure as devices change
- Failure model: when phone is disconnected, or even intermittently connected, applications don't work
- The portion of an app that makes sense to offload changes based on the network conn. to the cloud server

dynamic offloading

Application Partitioning

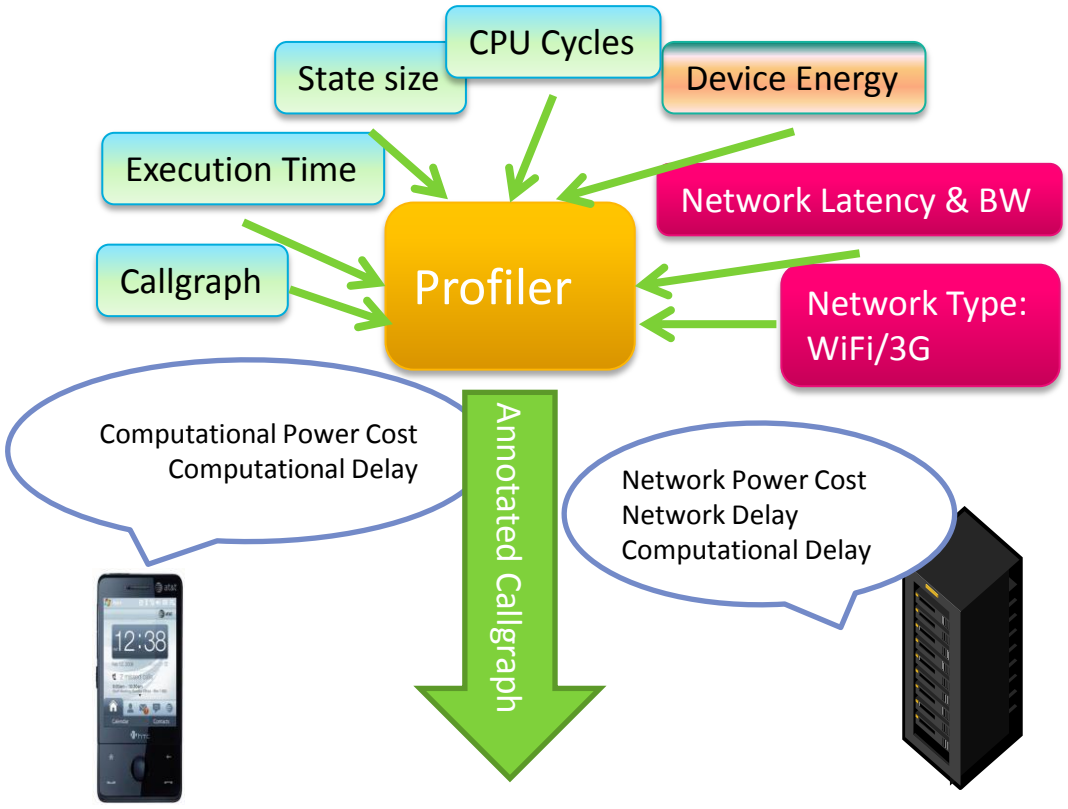


client/server split, can be extended to multiple tiers

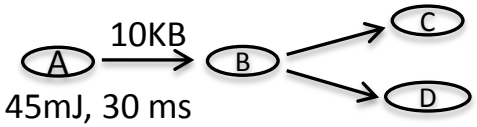
profiler and decision engine

Profiler:

Handles dynamics of devices, program behavior, and environment (Network, Server Load)



- Vertex: method annotated with computation energy and delay for execution
- Edge: method invocation annotated with total state transferred



Decision Engine:

Partition A Running App

We use an Integer Linear Program (ILP) to optimize for performance, energy, or other metrics…

Example – Maximize:

$$\sum_{v \epsilon V} (I_v \times E_v) \quad - \quad \sum_{(u,v) \epsilon E} (|I_u - I_v| \times C_{u,v})$$

energy saved — cost of offload

Such that:

$$\sum_{v \epsilon V} (I_v \times T_v) \quad + \quad \sum_{(u,v) \epsilon E} (|I_u - I_v| \times B_{u,v}) <= \text{Lat.}$$

execution time — time to offload

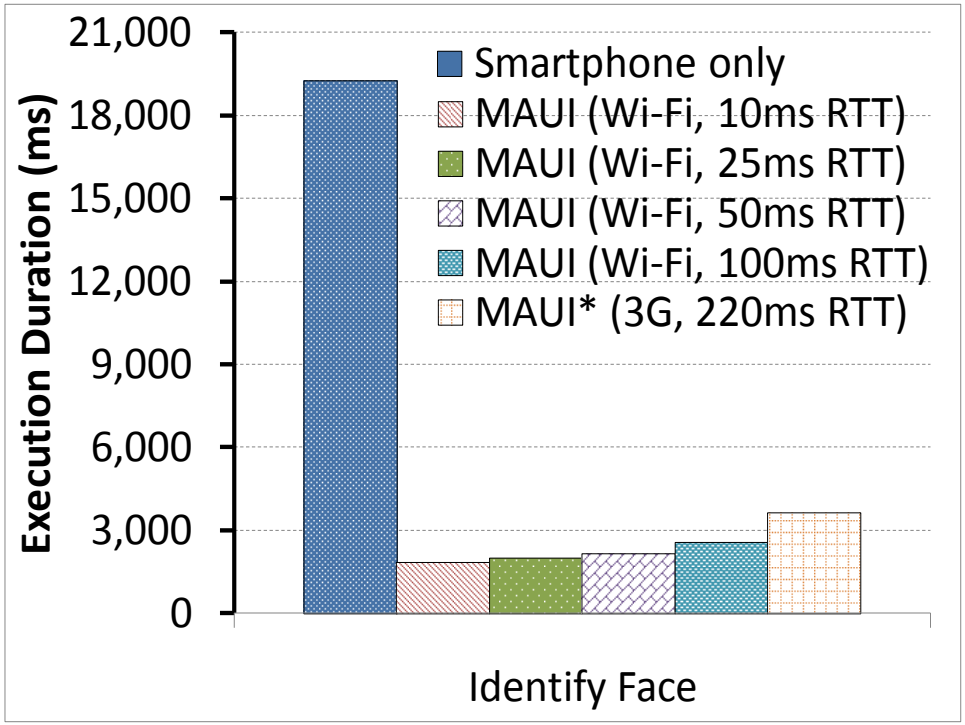
and

$I_v <= R_v$ for all $v \epsilon V$

performance and energy benefits

Performance Benefits:

Memory Assistant Face recognizer:



Face recognition becomes “interactive” w/ offload

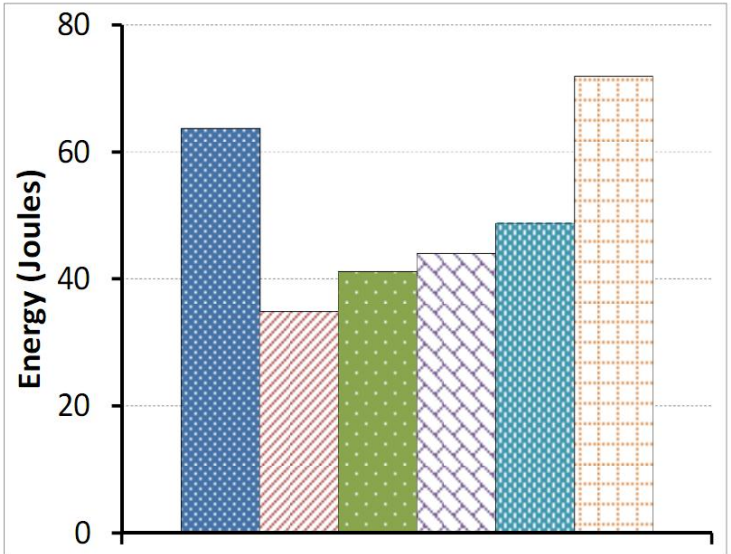
Energy Benefits:

Interactive arcade game w/physics engine:

Energy measurements from hardware power monitor

Arcade game benefits:

- Up to double the frame rate (6 -> 13 fps)
- Up to 40% energy reduction



alternate programming models

- MAUI: exploits .NET framework to dynamic partitioning & offload method execution **[Microsoft, MobiSys'10]**
- Odessa: creates a data-flow graph to exploit parallelism **[USC, MobiSys 2011]**
- CloneCloud: supports existing applications, but requires tight synchronization between cloud and phone **[Intel, EuroSys 2011]**
- Orleans: a new programming model based on grains **[Socc'11]**

	MAUI	CloneCloud	Odessa	Orleans
Remote execution unit	Methods (RMI)	Threads	Tasks	Grains

summarizing

- code offload allows developers to bypass resource limitations of handheld devices

- with dynamic offload, programmers no longer worry about *where* their code runs

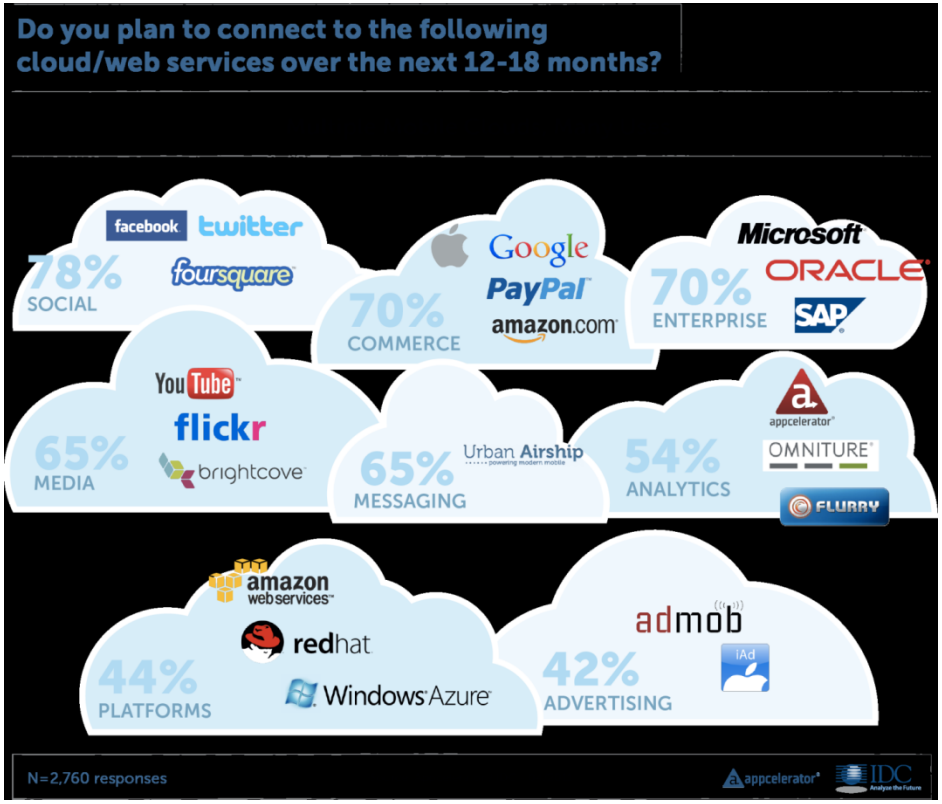
- leverage Microsoft's .NET runtime, Windows Phone OS, networking, Azure service, and Hyper-V security

Encourages developers to build applications they would never have considered possible.

cloud services

today: cloud offerings

focus is on providing infrastructure for storage and computation



Some heavy-duty web services available: email, search, etc.

state of art

Apple iCloud

- Store content in cloud and sync to all registered devices
- Hosted by Windows Azure and Amazon AWS
- iCloud storage APIs support third-party app document syncing

Amazon Silk

- Accelerates web access by learning user behavior then employing pre-caching
- Partitions work between local and AWS

OneLive

- Remote desktop, with fancy compression

future: the “service store”

… build world-class cloud services that enable application developers to easily realize the full potential of mobile computing

Examples:

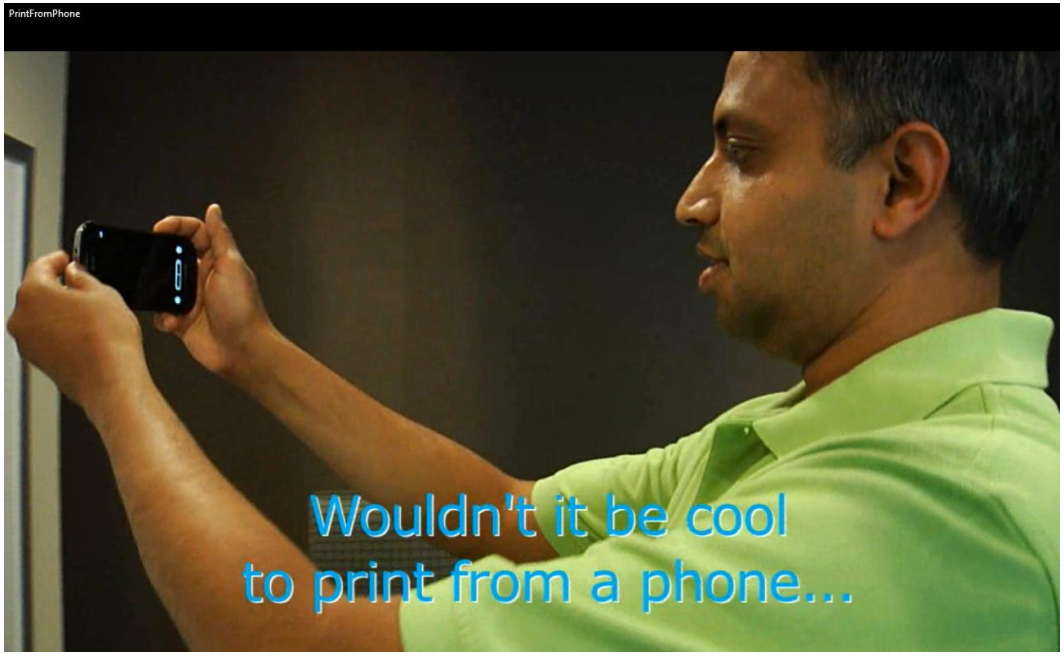
- Rendezvous: Lookup for relay endpoints
- Relay: Phone to phone data transfer
- Optical character recognition
- Speech-to-text, text-to-speech
- Face recognition, object recognition
- Multiplayer matchmaking
- Path prediction
- Social Mobile Sharing for ad hoc groups
- Speech Interface
- ….

Toolbox of services

sophisticated resource intensive algorithms running in the cloud typically CPU, memory & storage intensive battery and/or bandwidth hungry

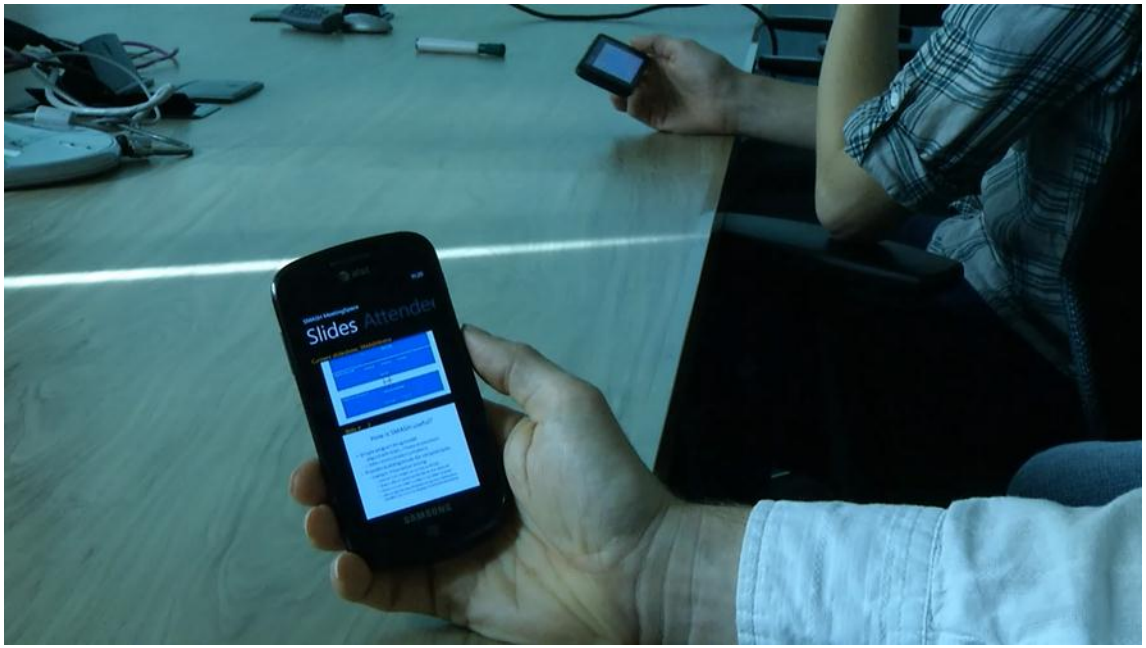
print from phone

example of using the OCR service



SMASH (relay service)

- Social mobile ad-hoc meeting support
- Built using relay & rendezvous



zero-effort payments

Bluetooth/BLE continuous scanning → Face recognition → Human assistance → Payment

TouchPoint device

FaceSDK

Cashier

matchmaking service
(for multi-player gaming)

multiplayer mobile gaming: challenge

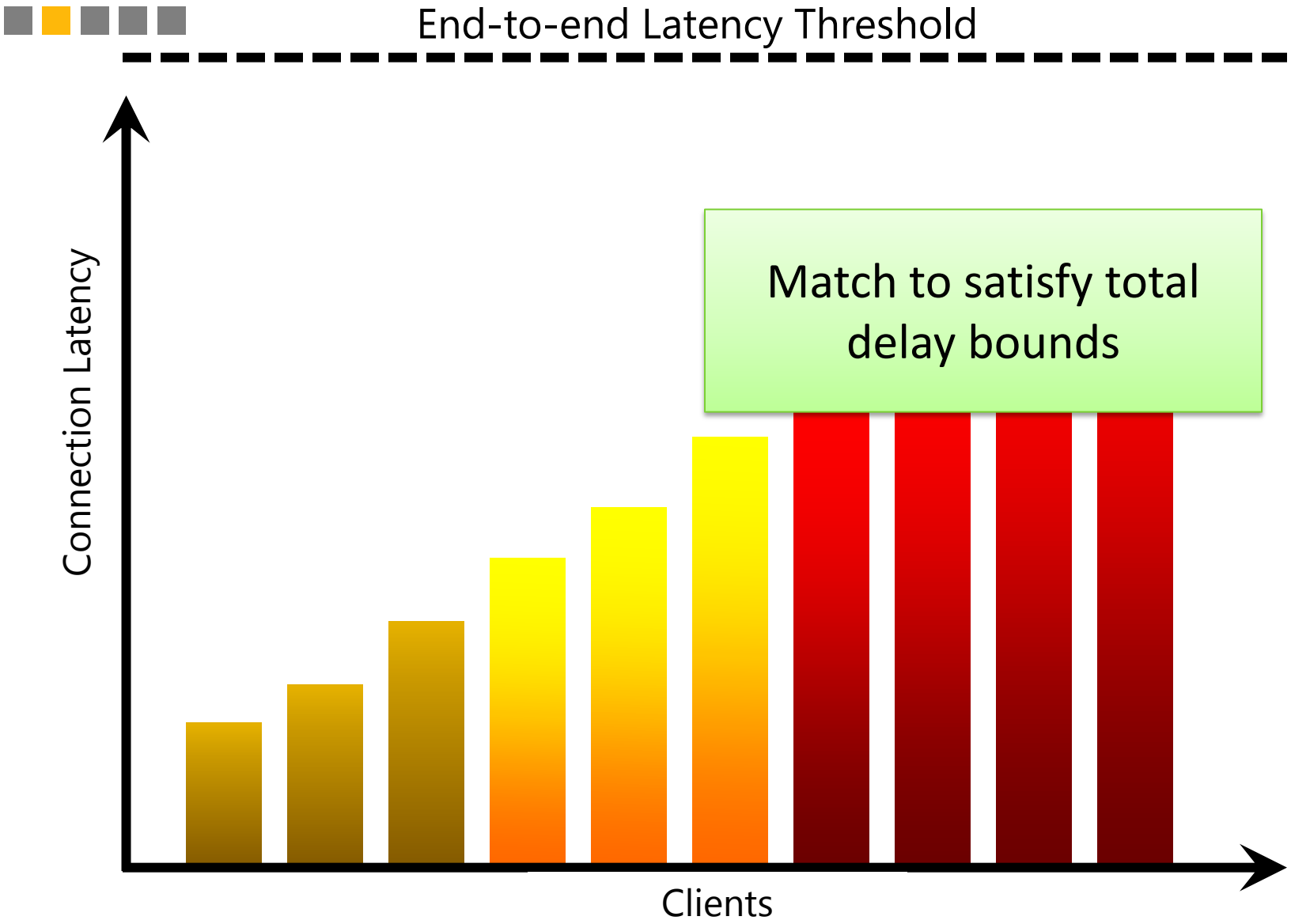
Bandwidth is fine: 250 kbps to host 16-player Halo 3 game

Delay bounds are much tighter

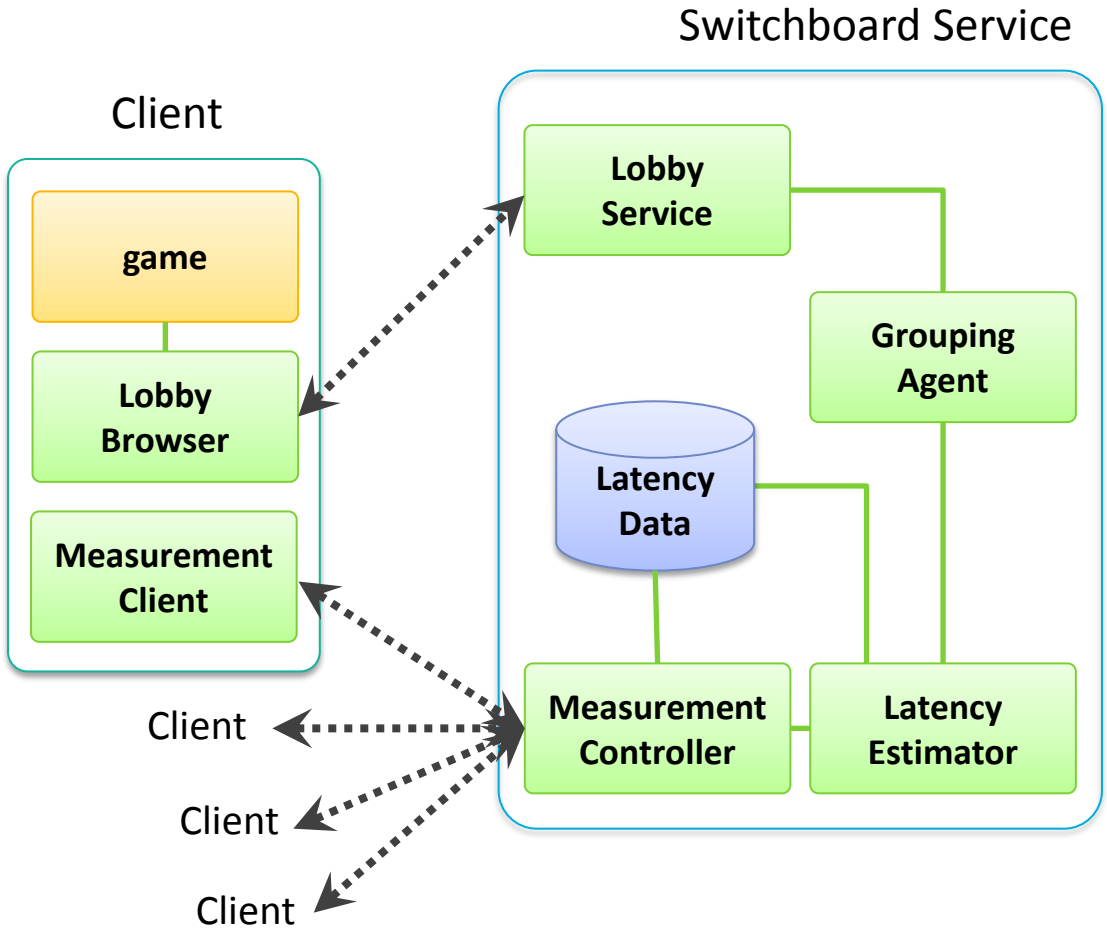
Challenge: find groups of peers than can play well together

Game Type	Latency Threshold
First-person, Racing	≈ 100 ms
Sports, Role-playing	≈ 500 ms
Real-time Strategy	≈ 1000 ms

the matchmaking problem



switchboard: matchmaker service



3G Measurement Study:

- Phone-to-phone latency stable over 15 minute intervals
- Can share latency profiles between phones using same cell tower

destination prediction service

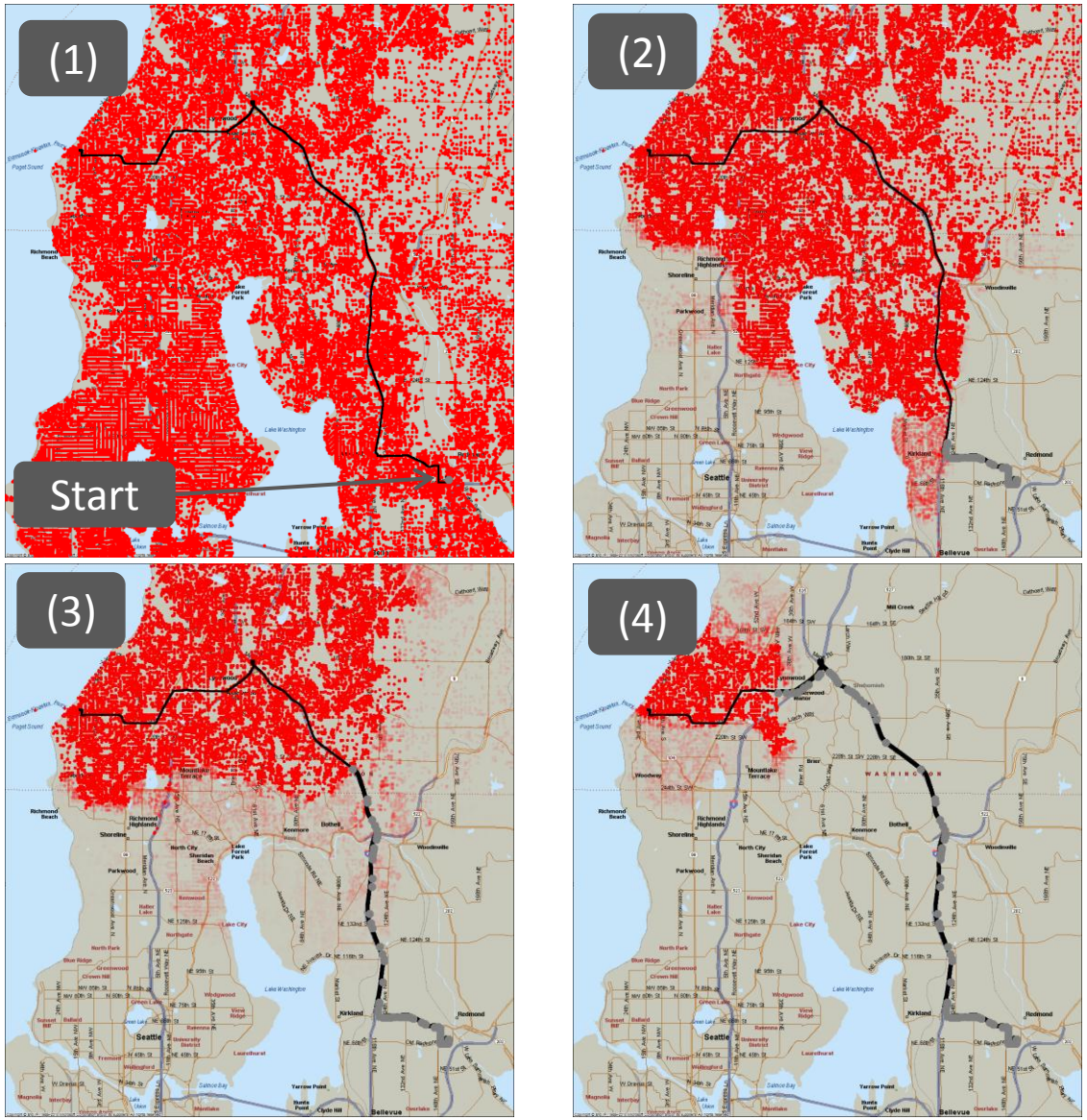
destination prediction service

Predict your destination as you drive

- Applications
  - Warn users of upcoming traffic incidents
  - Help find convenient stop (e.g. gas, coffee, food)
  - Target local search results to places ahead of you rather than behind you
  - Local ads for upcoming businesses

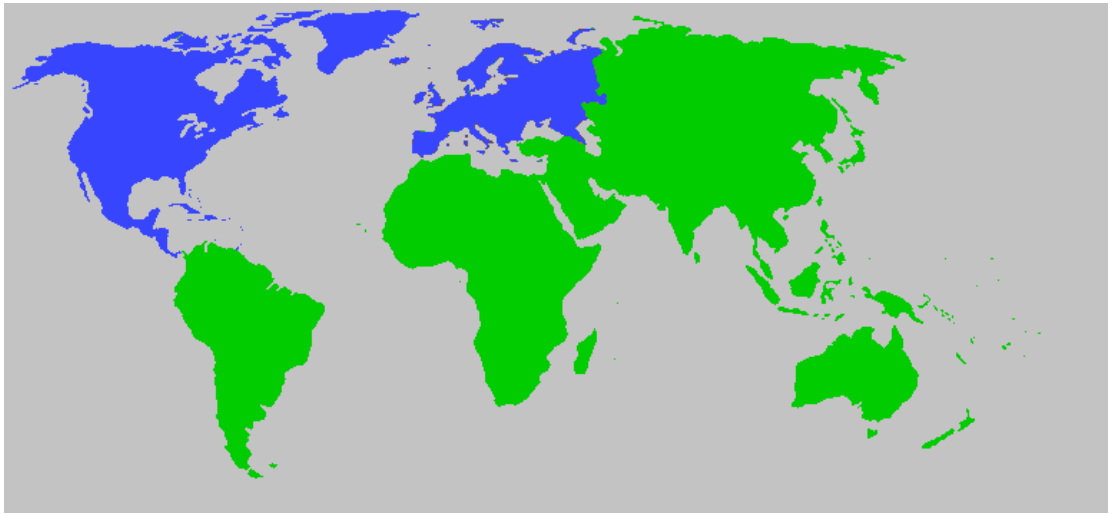
example trip

- Assumes driver takes (somewhat) efficient route to unknown destination
- Stores no GPS data, so privacy concerns reduced



algorithm & geographic Coverage

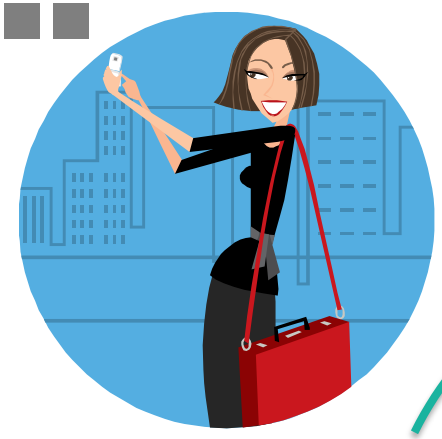
- Depends on driving time to ALL candidate destinations (“single-source shortest path”)
- Use PHAST algorithm from MSR SVC to do this really fast



- Prediction algorithm uses road network
- North America/Europe only for now

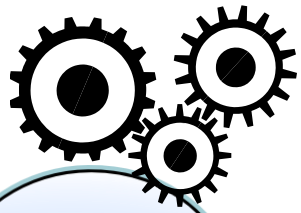
language modeling service

language modeling



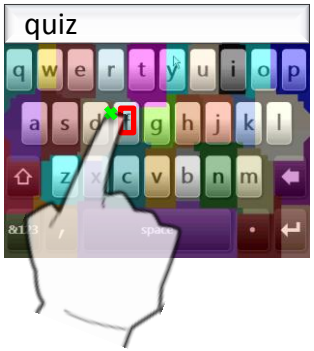
Data Collection

Processing



Cloud Services

Model Evaluation



- Hit-target resizing
- Spell-correction
- IME completions
- etc.

Deployment

OEM firmware

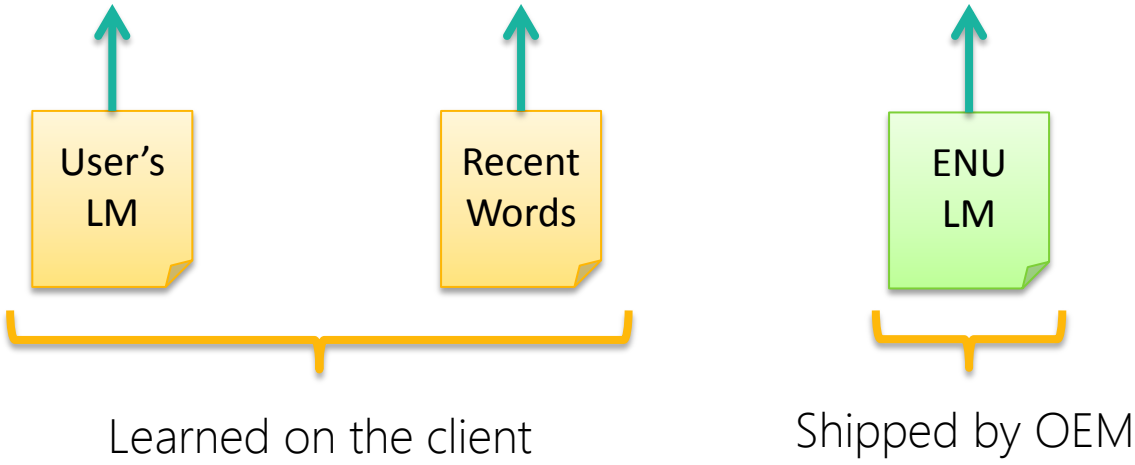
~~OTA Updates~~

user adaptation

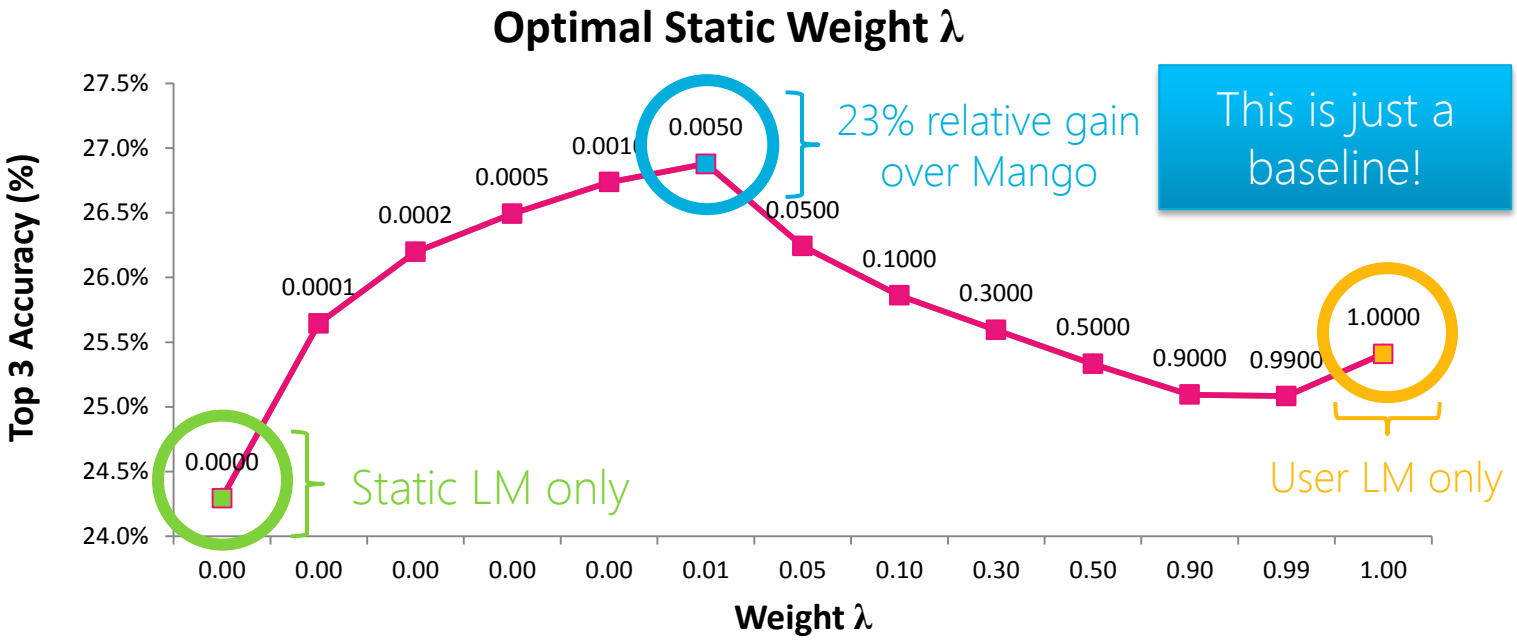
- Predicting users' language provides user delight
- To support user adaptation, we have been developing a dynamic interpolation framework:

$$p_{adapt}(w_i|w_{i-1}w_{i-2}) = \lambda_1 p_{user}(w_i|\dots) + \lambda_2 p_{recent}(w_i|\dots) + \lambda_3 p_{static}(w_i|\dots)$$

$$\sum_i \lambda_i = 1$$



Determining lamda



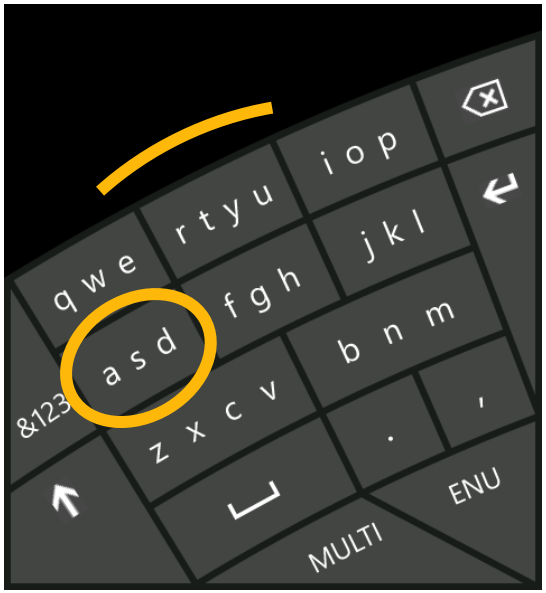
next word prediction + fixed candidates

DEMO

what else is baking?

- New Korean SIPs
  - No dominant SIP in Korea
  - Extending national SIPs for Apollo+
  - New eye's-free SIP

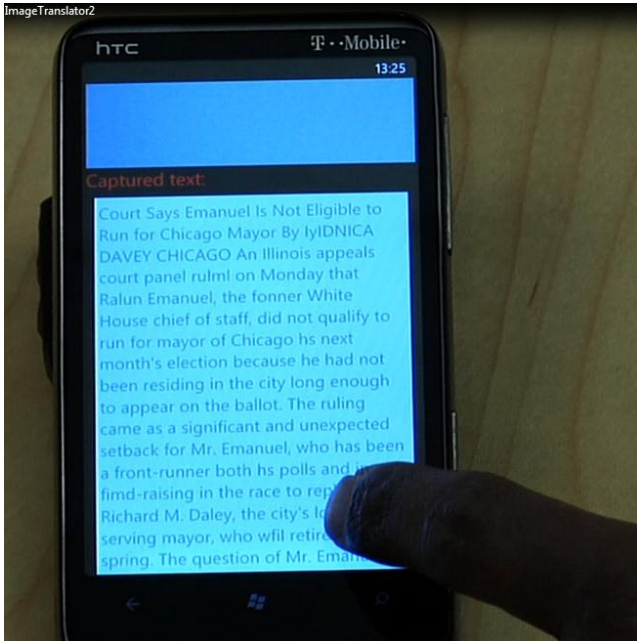
- Natural Arc
  - Ergonomically designed for 1-handed thumb usage
  - Keys cluster several letters
  - Leverages disambiguation



natural arc

DEMO

composing services



Glue that holds various cloud services together

- Connects services together & provides simple custom logic
- Eliminates multiple round-trips to the client

trying it in the real world...

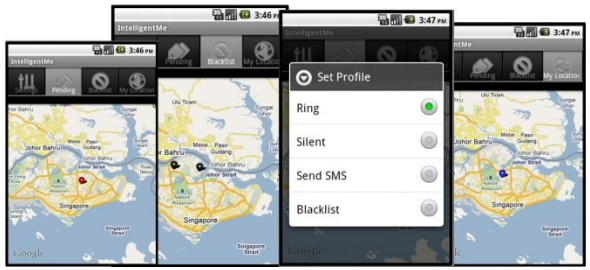
Project Hawaii

Unleashing the creative power of students by lowering barriers to writing mobile + cloud apps

Hawaii academic program

- Hawaii courses taught over last 2 years:
  - 61 universities, 915 students, > 100 successful app projects

student developed applications

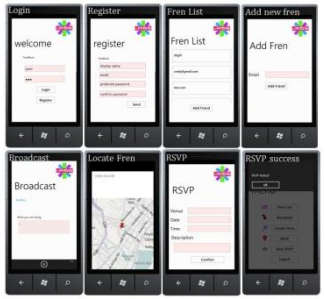


intelligentME

ReceiptManager

DaySaver

Network Forecaster

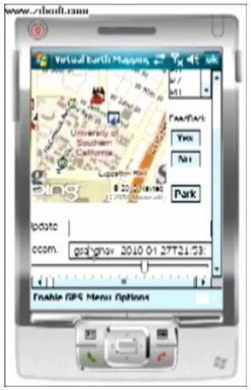


myFrens

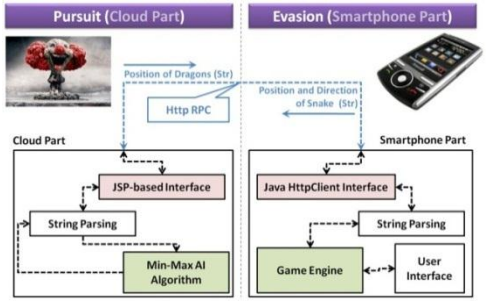
LunchBox

Flagged Down

MobiProg

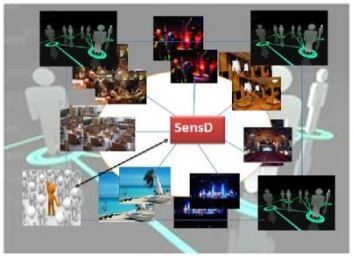


Parking Assistant



Snakes & Dragons

Activity Classification



SensD

Image Stitching

Microsoft Research
Project Hawaii

Sample Press Articles

web 2.0 SUMMIT

OCTOBER 17–19, 2011 • SAN FRANCISCO, CA

Home Register Conference Sponsors Video Press & Media Connect Your Account STAY CONNECTED

CAMPUS TECHNOLOGY

Home News Features Opinion

Taking Mobile Applications into the Cloud

Cloud-enabled mobile computing is at the intersection of two of today's hottest areas in IT, coupling resource-starved mobile phones with the resource-rich cloud.

ZDNet 20th Anniversary

Microsoft Research delivers cloud development kit for Windows Phone

Microsoft's new Hawaiian codenames are all about mobile

By Mary Jo Foley | July 9, 2010, 12:04pm PDT

WPowerMUser.com

Developers: Windows Phone 7 + Cloud Services SDK By MSR

January 27, 2011 | By Pradeep

cnet Japan

All About Microsoft

マイクロソフト、クラウド対応モバイルアプリ研究プロジェクト「Hawaii」

hawaii services roadmap

Deployed	Prototyping	Design Stage
Rendezvous: Lookup for relay endpoints	Mobile Game Matchmaking	NLify
Relay: Phone to phone data transfer	Trajectory Prediction	Location Sharing
Optical Character Recognition	SMASH social, mobile sharing for ad-hoc meetings	Generic Machine Learning
Speech2Text	Face recognition	GeoFencing
Key-value store	Service Composition	Generic Offload Services

… plus existing services: WP Location, WP Notification, Bing Maps, Bing Translation

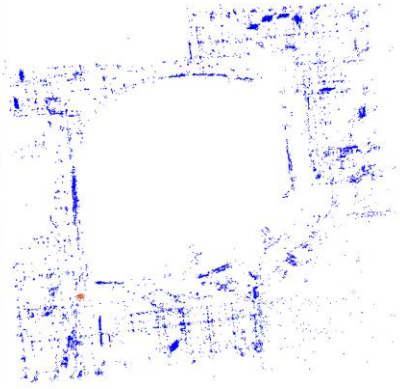
All services are integrated with Azure MarketPlace

phones that see

Looxcie, Inc

who?

what?

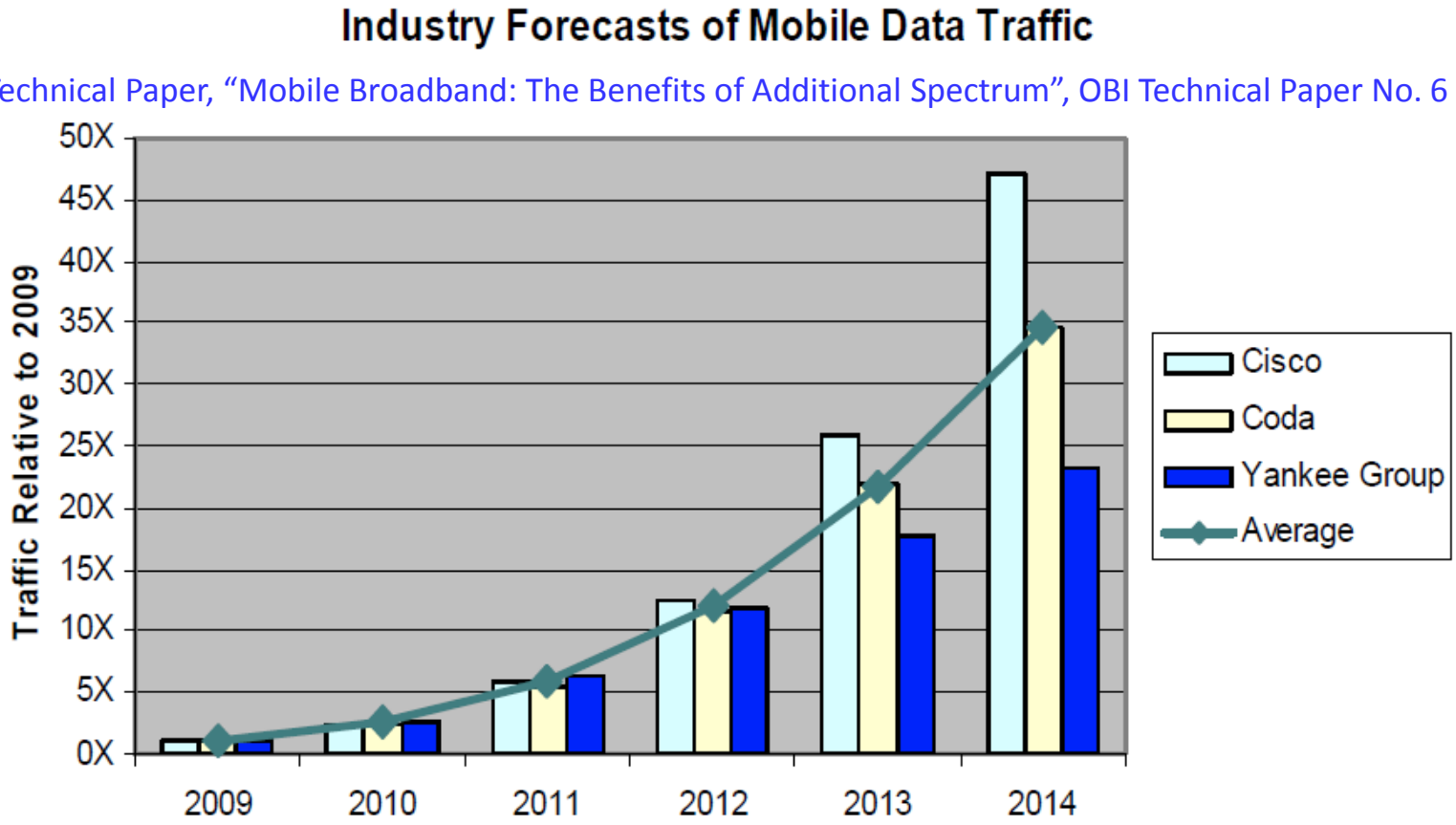


where?

Video credits:
Matthai Philipose, MCRC
Intel Labs

what about connectivity?

bandwidth scarcity: demand continues to go up

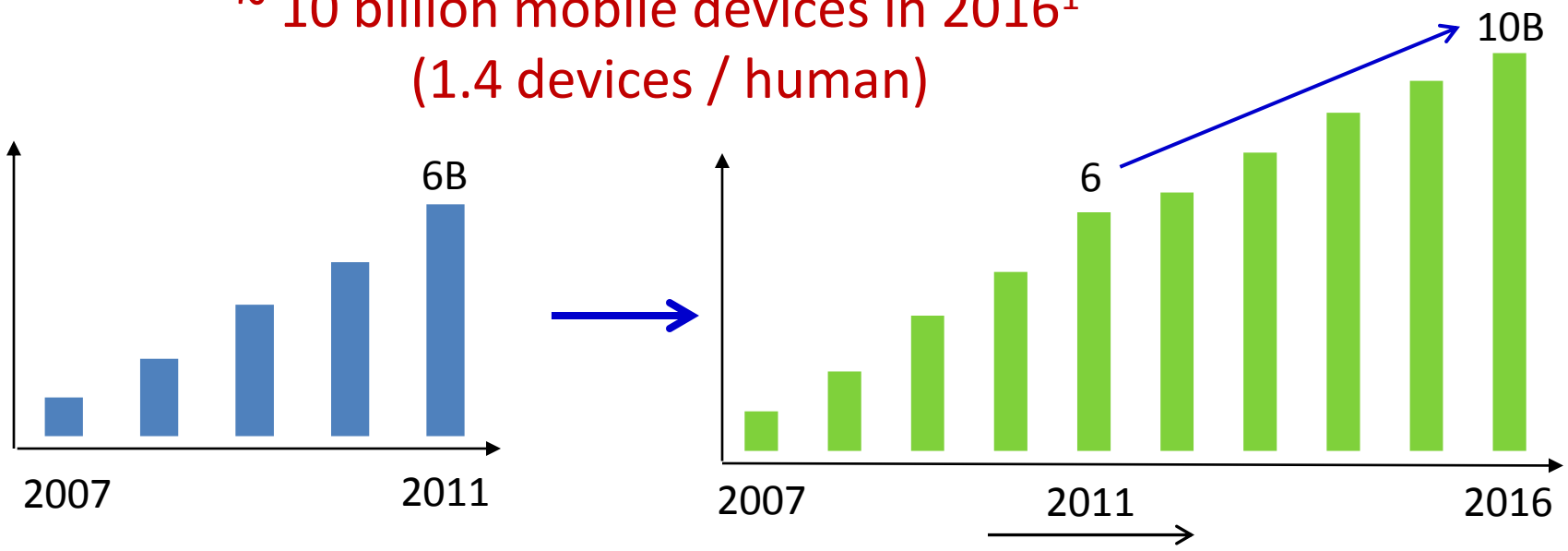


FCC, Staff Technical Paper, "Mobile Broadband: The Benefits of Additional Spectrum", OBI Technical Paper No. 6 (Oct. 2010),

AT&T's mobile data traffic has experienced a fifty-fold increase over a three year period

mobile connected devices: growth

~ 10 billion mobile devices in 2016[1]
(1.4 devices / human)



2011-2016 ~ 18X growth in mobile data traffic[2]
(~ 10 exabytes / month)

Source: (1) GSMA; (2) Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011–2016

from vision to strategy

getting to 100K+ cloud enabled apps.

- focus on the developer
  - → computational offload
  - → Resource intensive cloud services
  - → multi-device programming

- focus on ubiquitous connectivity to the cloud
  - → cut down latency & mitigate bandwidth scarcity (e.g. cloudlets)
  - → opportunistic networking (e.g. White spaces)

TestMyNet

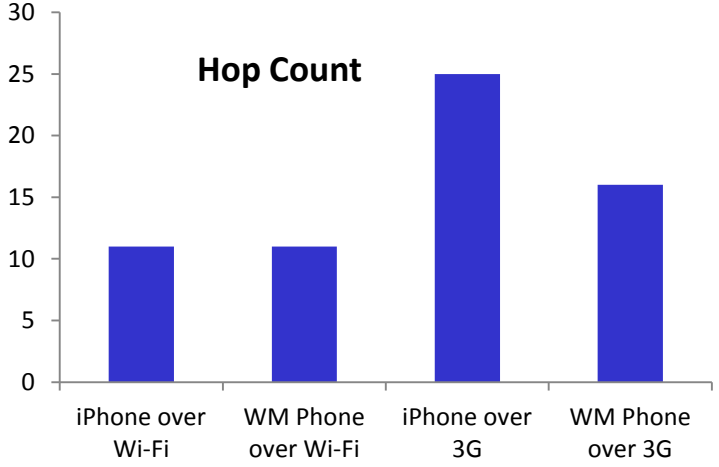
Available on Windows Phone Maretplace
65 Reviews, averages review rating of 4.75/5 stars

latency

iPhone via Wi-Fi : 11 hop

Wi-Fi -> 209.85.225.99

1. (10.0.2.1) 8.513 ms 8.223 ms 9.365 ms
2. (141.212.111.1) 0.913 ms 0.606 ms 0.399 ms
3. (192.122.183.41) 11.381 ms 6.054 ms 5.975 ms
4. (192.12.80.69) 7.038 ms 7.353 ms 7.026 ms
5. (198.108.23.12) 12.525 ms 13.027 ms 12.619 ms
6. (198.110.131.78) 12.715 ms 9.424 ms 9.315 ms
7. (216.239.48.154) 9.974 ms (209.85.250.237) 10.295 ms (216.239.48.154) 9.405 ms
8. (72.14.232.141) 19.308 ms 22.249 ms 23.312 ms
9. (209.85.241.35) 32.987 ms 22.708 ms (209.85.241.27) 124.588 ms
10. (72.14.239.18) 22.256 ms (209.85.248.106) 29.154 ms (209.85.248.102) 21.635 ms
11. (209.85.225.99) 19.973 ms 21.930 ms 21.656 ms



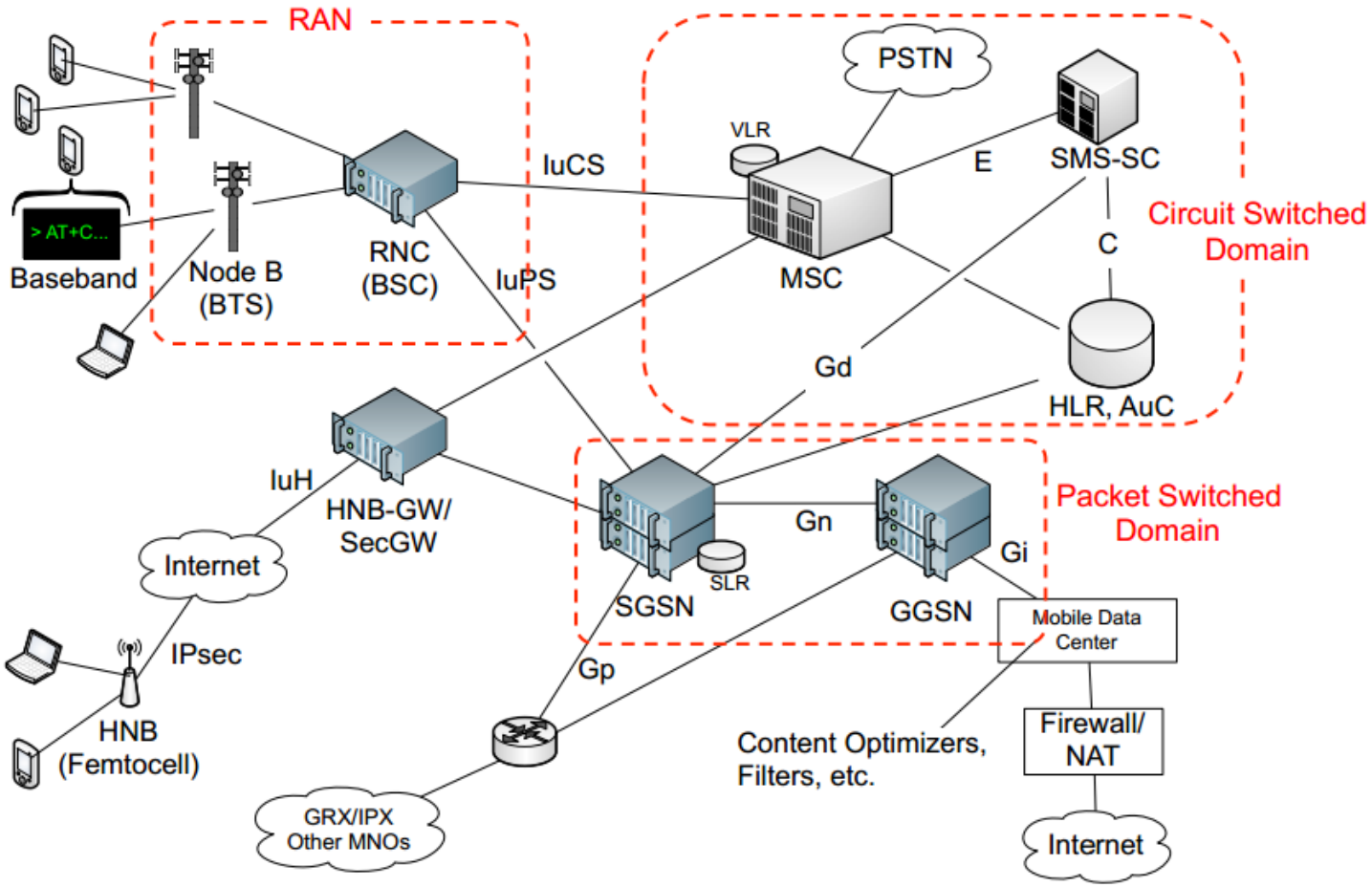
iPhone via 3G : 25 hop

3G -> 209.85.225.99

1. * * *
2. (172.26.248.2) 414.197 ms 698.485 ms 539.776 ms
3. (172.16.7.82) 1029.853 ms 719.595 ms 509.750 ms
4. (10.251.11.23) 689.837 ms 669.340 ms 689.739 ms
5. (10.251.10.2) 509.781 ms 729.746 ms 679.787 ms
6. (10.252.1.7) 719.652 ms 760.612 ms 788.914 ms
7. (209.183.48.2) 689.834 ms 599.675 ms 559.694 ms
8. (172.16.0.66) 539.712 ms 809.954 ms 689.547 ms
9. (12.88.242.189) 589.857 ms 1129.848 ms 709.784 ms
10. (12.122.138.38) 589.699 ms 1009.723 ms 769.808 ms
11. (12.122.138.21) 669.690 ms 529.758 ms 699.965 ms
12. (192.205.35.222) 699.569 ms 979.769 ms 1489.869 ms
13. (4.68.19.190) 699.435 ms (4.68.19.126) 559.875 ms (4.68.19.62) 499.598
14. (4.69.136.149) 889.946 ms (4.69.136.141) 879.443 ms (4.69.136.145) 469.601 ms
15. (4.69.132.105) 559.716 ms 539.754 ms 1219.982 ms
16. (4.69.132.38) 719.700 ms 659.613 ms 539.695 ms
17. (4.69.132.62) 549.752 ms 549.640 ms 800.128 ms
18. (4.69.132.114) 669.729 ms (4.69.140.189) 769.711 ms 959.663 ms
19. (4.69.140.193) 959.735 ms 979.674 ms 849.886 ms
20. (4.68.101.34) 649.609 ms 659.767 ms (4.68.101.98) 1119.996 ms
21. (4.79.208.18) 669.405 ms 629.574 ms (209.85.240.158) 1200.039 ms
22. (209.85.240.158) 769.538 ms (72.14.232.141) 729.505 ms (209.85.241.22) 719.715 ms
23. (209.85.241.22) 769.665 ms (209.85.241.35) 769.880 ms 859.536 ms
24. (209.85.241.29) 589.710 ms (66.249.95.138) 789.762 ms (209.85.248.106) 913.287 ms
25. (209.85.225.99) 716.000 ms (66.249.95.138) 1039.963 ms (72.14.239.18) 899.607 ms

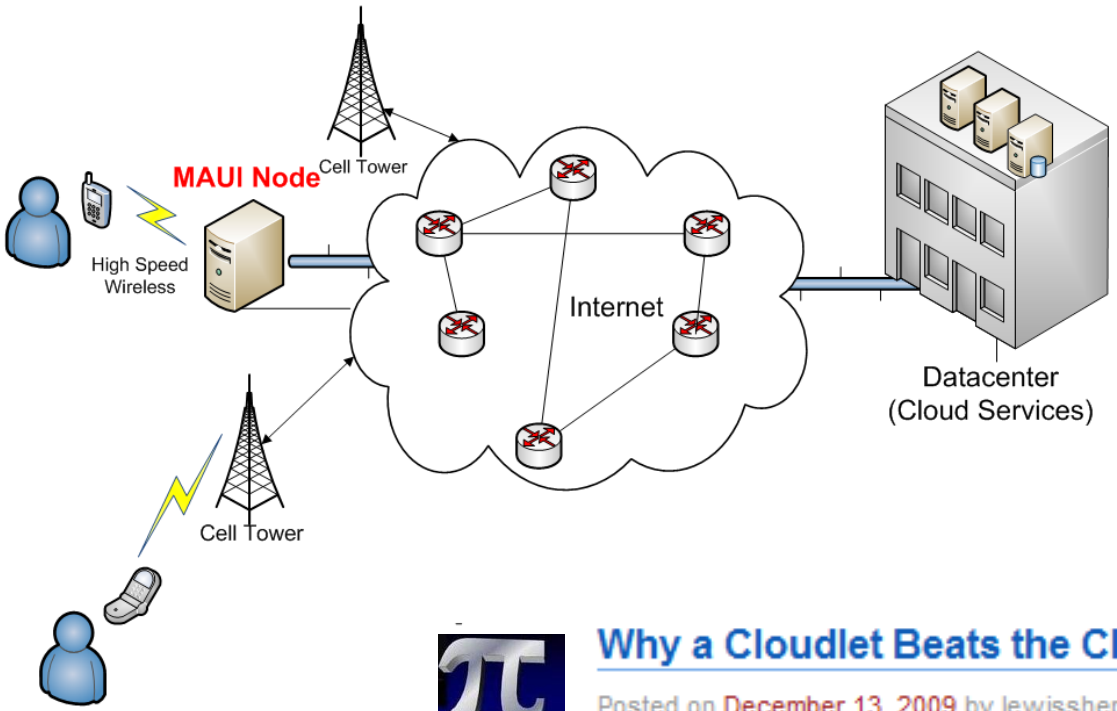
traceroute to 209.85.225.99 (one of the server IPs of www.google.com)

heavyweight architecture



cloudlets: defined

a resource rich infra-structure computing device with high-speed Internet connectivity to the cloud that a mobile device can use to augment its capabilities and enable applications that were previously not possible



Why a Cloudlet Beats the Cloud for Mobile Apps

Posted on December 13, 2009 by lewisshepherd

sample deployment scenario

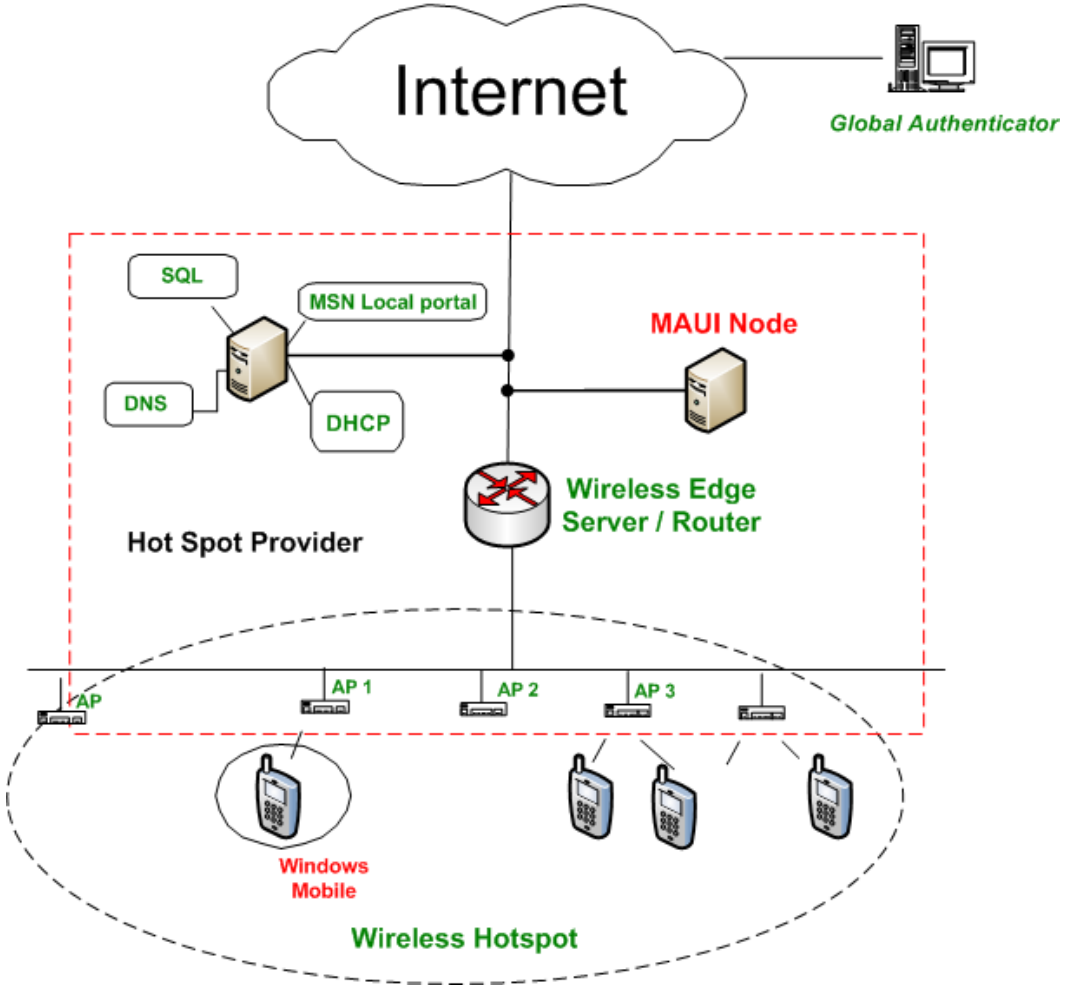
augment Wi-Fi hot spots with cloudlets
(in publics spaces & enterprise networks)

advantages

- does not use cellular spectrum
- short round-trip-times between mobile & cloud(let)
- optimal performance

research challenges

- Offload framework
- caching
- security & privacy



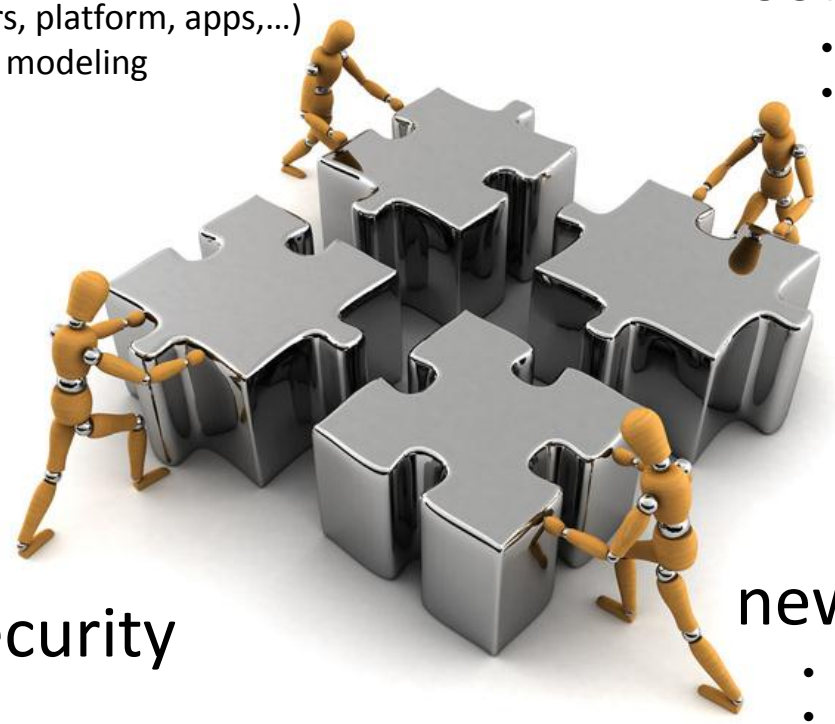
conclusion: highly interdisciplinary field

machine learning

- big data (sensors, platform, apps,…)
- predictions and modeling

software engineering

- empirical software eng.
- program analysis



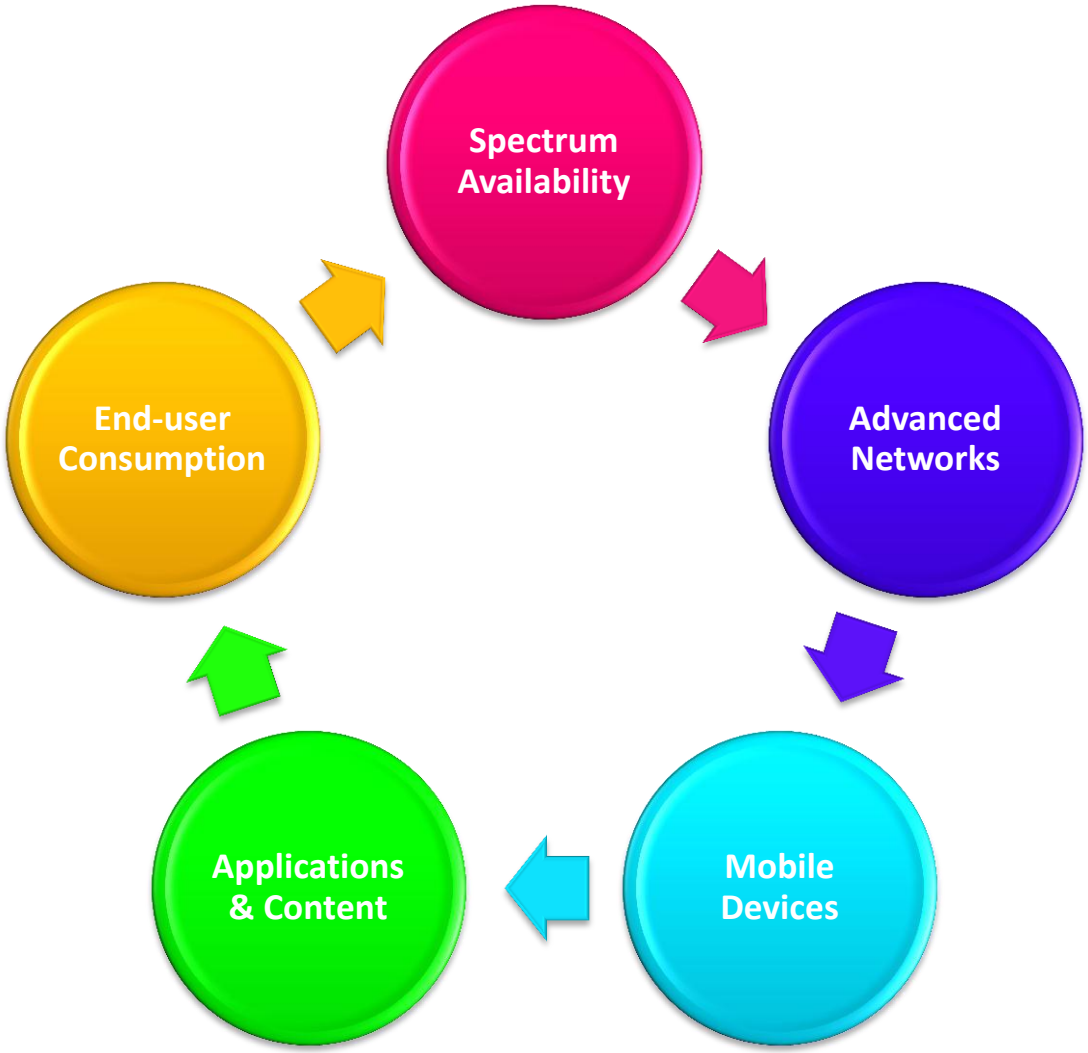
systems & security

- building to scale
- sensor systems
- energy management

new user experiences

- gestures, speech
- context awareness, social computing
- hardware accessories

mobile computing: virtuous cycle of innovation



Courtesy: Ralph de la Vega, President CEO AT&T Mobility

a bright future

Cloud

offloading + services

+

Rich Devices

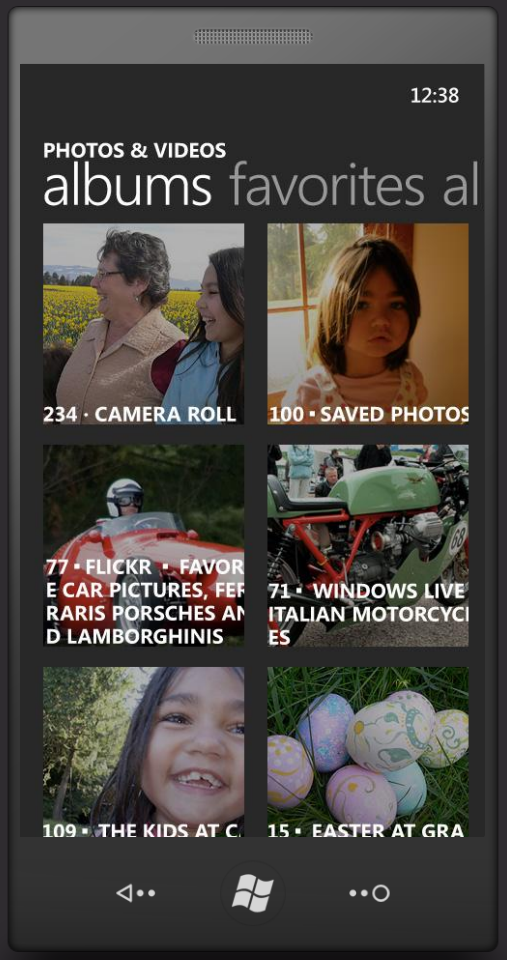
Rich Connectivity

=

plethora of enterprise class mobile computing apps


XBOX LIVE
MONOPOLY
NETFLIX
ebay
IMDb
XBOX LIVE
XBOX LIVE
The Revenants
REALTOR.com®
SLACKER
SLACKER RADIO
XBOX LIVE
XBOX LIVE
XBOX LIVE
Extras
XBOX LIVE
GUITAR HERO 5 MOBILE
AP
XBOX LIVE
HEXIC RUSH
XBOX LIVE
loopt
XBOX LIVE
Flowerz
XBOX LIVE
HALO WAYPOINT
photobucket
yelp
YouTube
amazonkindle
XBOX LIVE
SHOWTIME

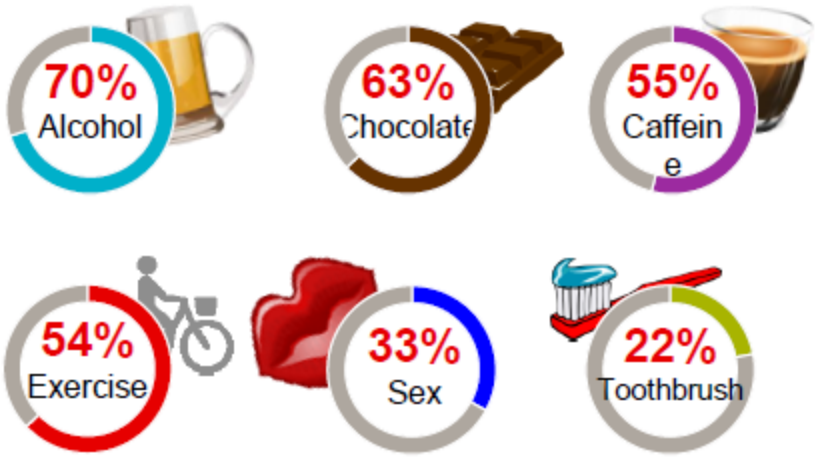
Thanks!



mobile computing

massive dependency / addition

Would rather give up[1]



57% use email

53% browse the web

38% social networking

30% download content

25% upload content

20% stream content

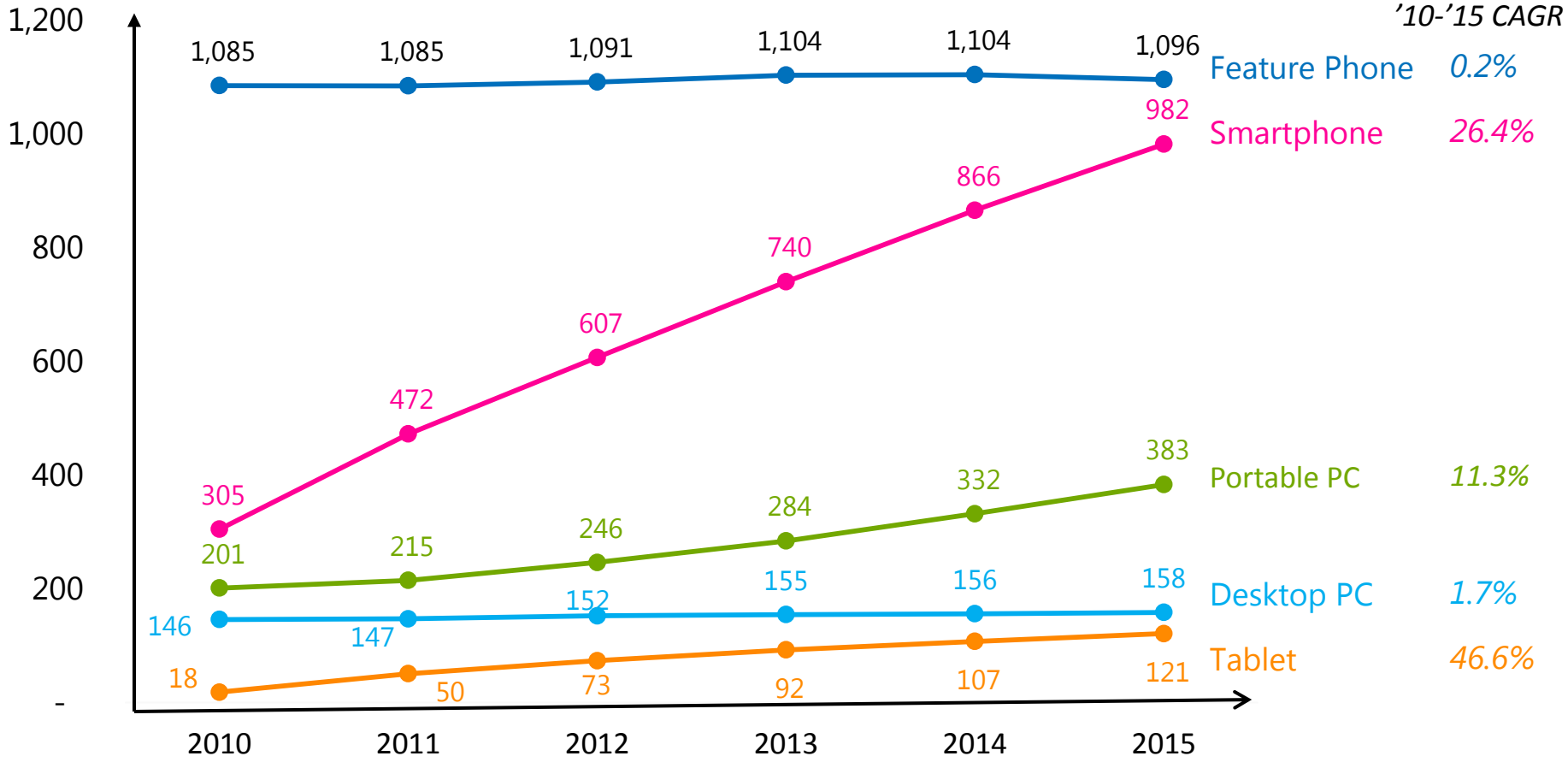
1. Telenav, US survey July 2011

 All other data: Vodafone in-house research

comparing growth trends

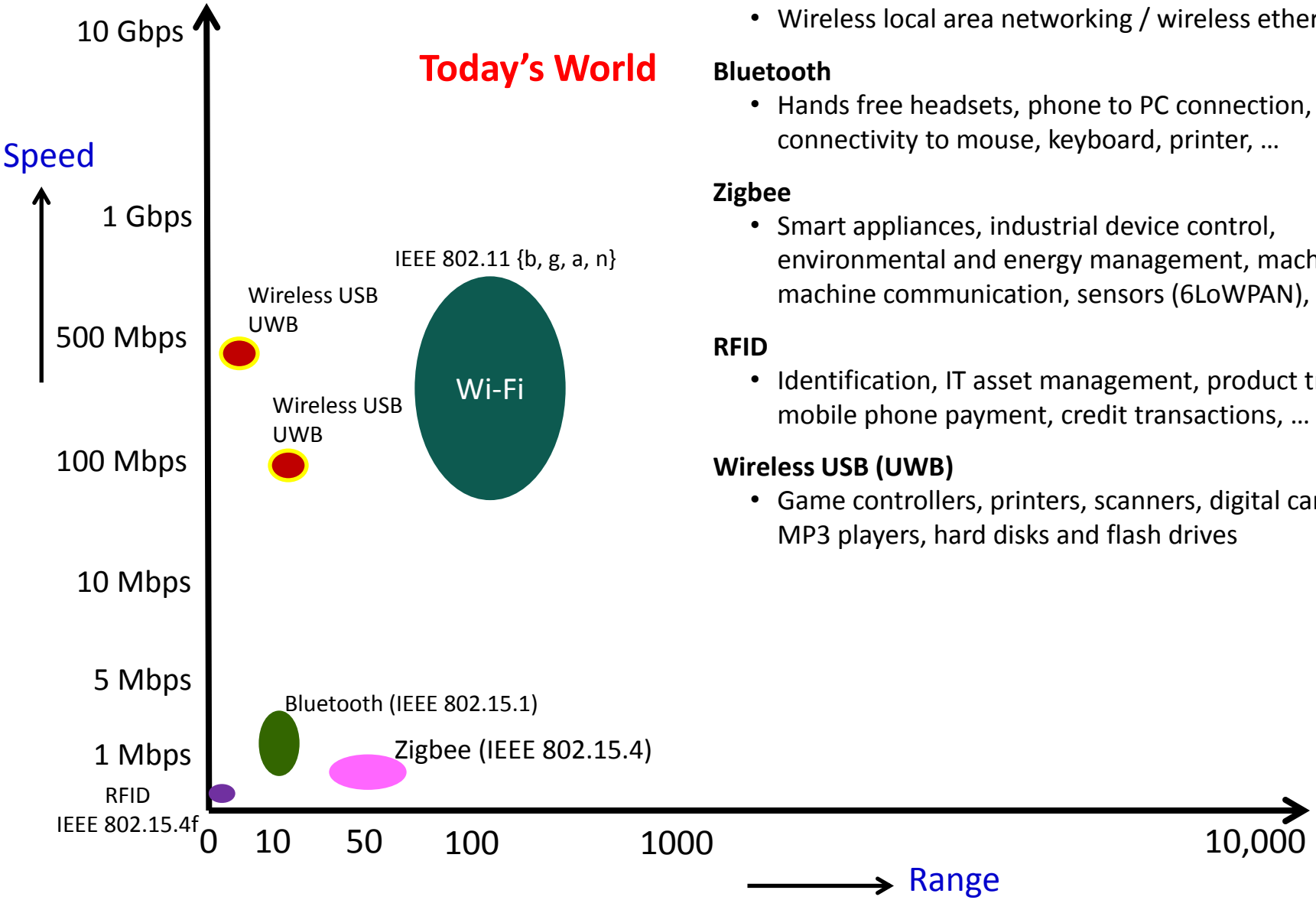
2010 – 2015, million units

SmartPhones outsell laptops & desktops



Sources: BLS, IDC, CEA, Accenture analysis

connectivity options over unlicensed frequencies



Wi-Fi

- Wireless local area networking / wireless ethernet

Bluetooth

- Hands free headsets, phone to PC connection, ad hoc connectivity to mouse, keyboard, printer, …

Zigbee

- Smart appliances, industrial device control, environmental and energy management, machine-to-machine communication, sensors (6LoWPAN), …

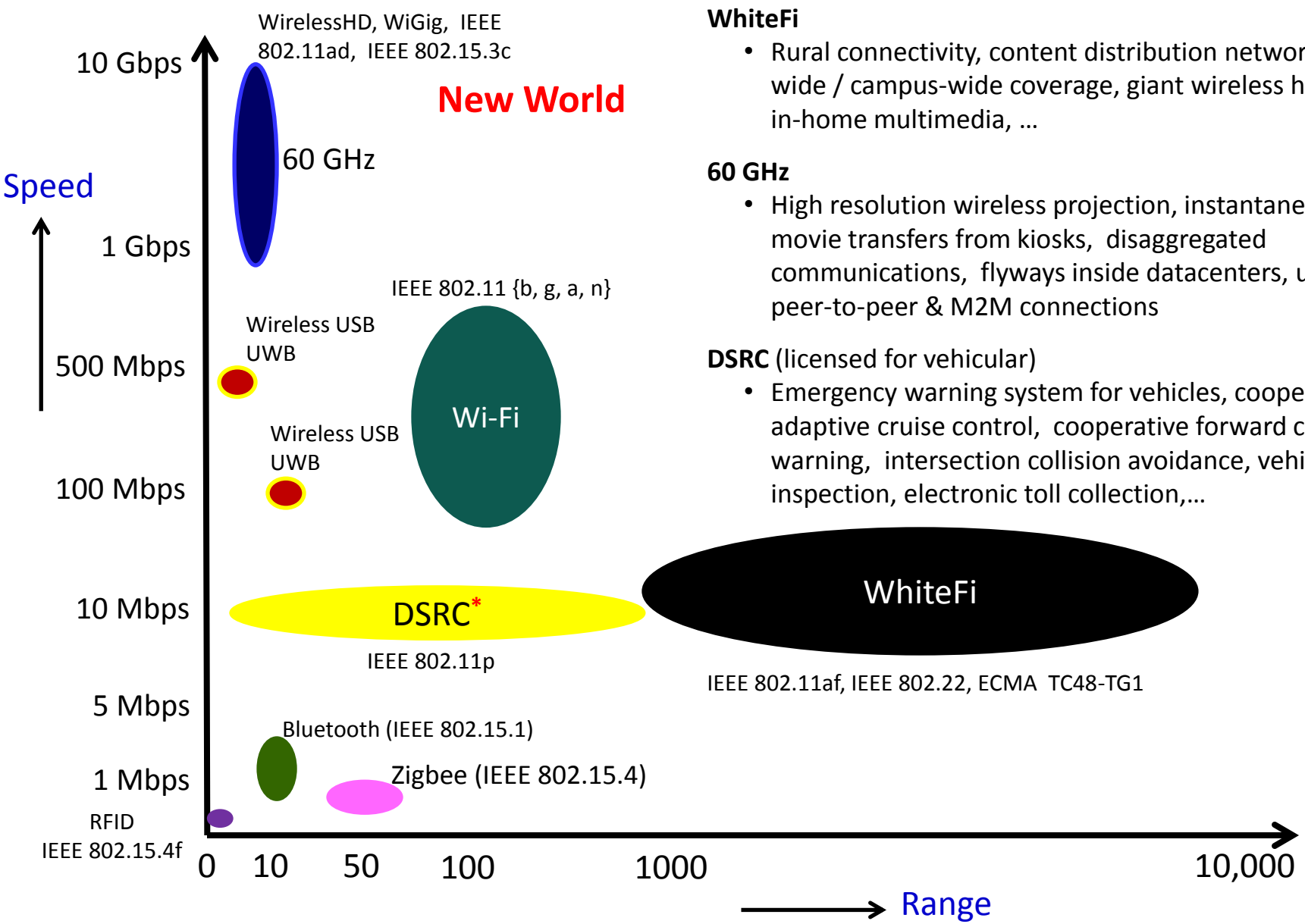
RFID

- Identification, IT asset management, product tracking, mobile phone payment, credit transactions, …

Wireless USB (UWB)

- Game controllers, printers, scanners, digital cameras, MP3 players, hard disks and flash drives

connectivity options over unlicensed frequencies



WhiteFi

- Rural connectivity, content distribution networks, city-wide / campus-wide coverage, giant wireless hotspots, in-home multimedia, …

60 GHz

- High resolution wireless projection, instantaneous movie transfers from kiosks, disaggregated communications, flyways inside datacenters, ultra-fast peer-to-peer & M2M connections

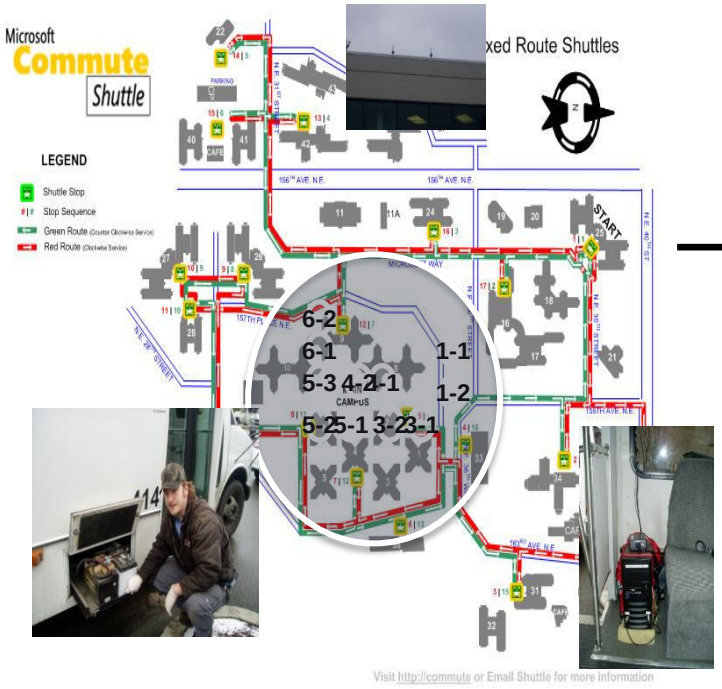
DSRC (licensed for vehicular)

- Emergency warning system for vehicles, cooperative adaptive cruise control, cooperative forward collision warning, intersection collision avoidance, vehicle inspection, electronic toll collection,…

the world's first urban white space network

Oct. 16, 2009

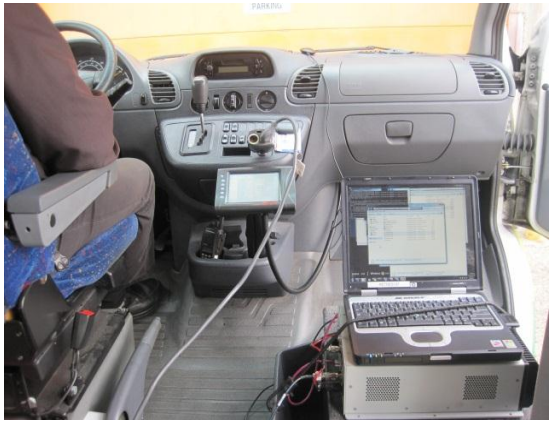
A giant white space hot-spot network on Microsoft campus



Accessing from the office

WS Antenna on Bldg 42

WS antenna on MS Shuttle



Accessing from inside a MS Shuttle

Victory!

MAY 11, 2010, 6:59 P.M. ET

FCC Officials Visit Microsoft To Examine Experimental Network

Aug 14, 2010

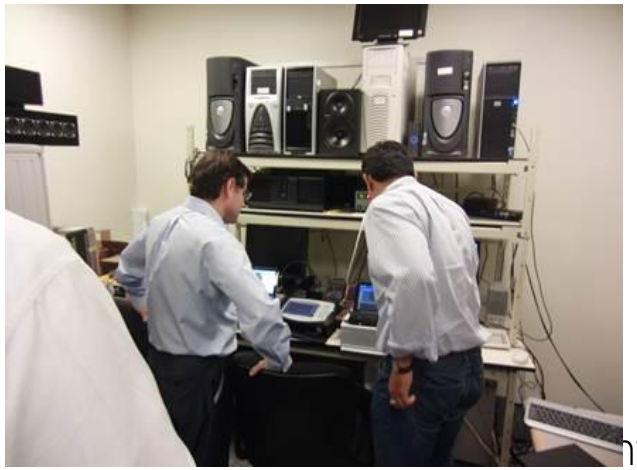
The New York Times

F.C.C. Opens Unused TV Airwaves to Broadband

By EDWARD WYATT
Published: September 23, 2010

Chairman Genachowski & Microsoft's CTO Craig Mundie, August 14, 2010

Chairman Genachowski and FCC Managing Director Steven VanRoekel Climb aboard the MS Shuttle to look at our WhiteFi Network



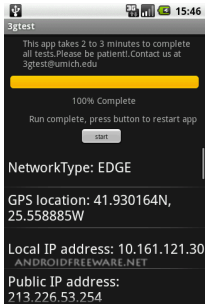
FCC Chairman Genachowski looks at our wireless Microphone demo In Bldg. 99, Anechoic Chamber (Room 1651)

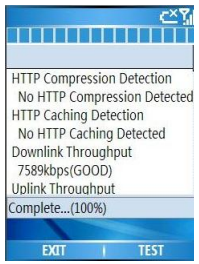
conclusion: integrating business & consumer needs

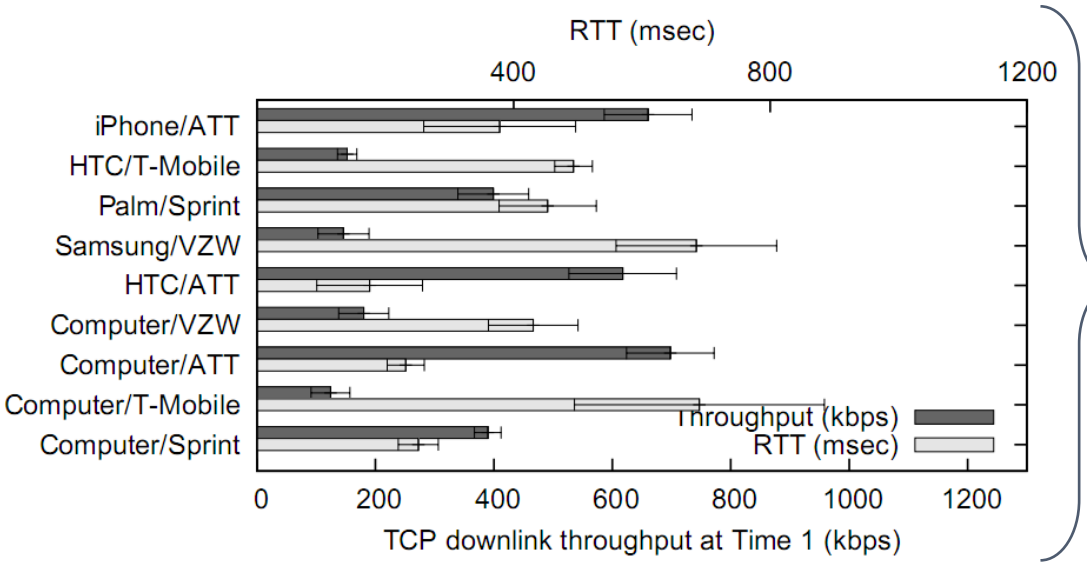
- compelling end-user experiences
- value to the organization
- captivating applications for customers and employees

cloud computing has its challenges

End-to-end latency hurts interaction quality (crisp interaction essential for low demand on human attention)







High loss rate & low throughput severely limits the scope of cloud services

http://www.eecs.umich.edu/3gtest

putting the cloud in the palm of your hands

Victor Bahl

5.23.2012

from vision to strategy

getting to 100K+ cloud enabled apps.

- focus on the developer
  - → programming support for computational offload
  - → Resource intensive cloud services and their composition
  - → cloud supported multi-device programming

- focus on ubiquitous connectivity to the cloud
  - → cut down latency & mitigate bandwidth scarcity (e.g. cloudlets)
  - → opportunistic networking (e.g. White spaces)