TechFest
the & in R&D
Microsoft Research

Microsoft Research

Microsoft

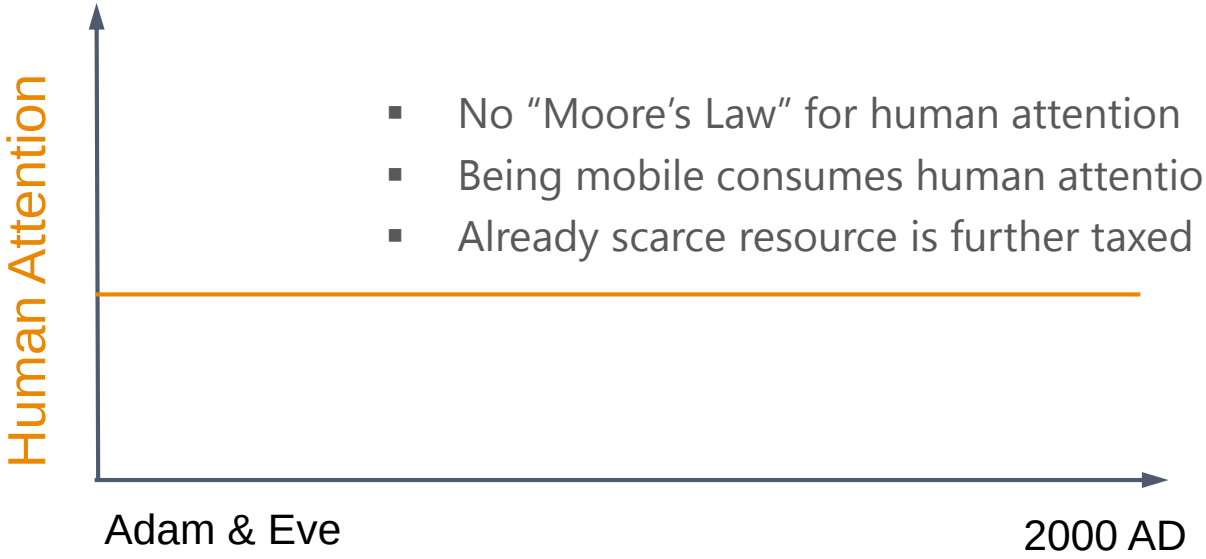
cloudlets for mobile computing

Victor Bahl

June 27, 2014

Courtesy. M. Satya, CMU

why resource poverty hurts



- No "Moore's Law" for human attention
- Being mobile consumes human attention
- Already scarce resource is further taxed by resource poverty

Reduce demand on human attention

- Software computing demands not rigidly constrained
- Many "expensive" techniques become a lot more useable when mobile

Some examples

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

…

Clever exploitation needed to deliver these benefits

Vastly superior mobile user experience

Microsoft

latency matters!

"Being fast really matters…half a second delay caused a 20% drop in traffic. and it killed user satisfaction"

- Marissa Mayer @ Web 2.0 (2008)

"…a 400 millisecond delay resulted in a -0.59% change in searches/user",

[i.e. Google *would lose 8 million searches per day - they'd serve up many millions fewer online adverts*]

- Jake Brutlag, Google Search (2009)

"…for Amazon every 100 ms increase in load times decreased sales with 1%"

- Andy King, book author

"…when 50% of traffic was redirected to our edges preliminary results showed a 5.9% increase in click-thru rates"

- Andy Lientz, Partner GPM, BingEdge (2013)

(also,)

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

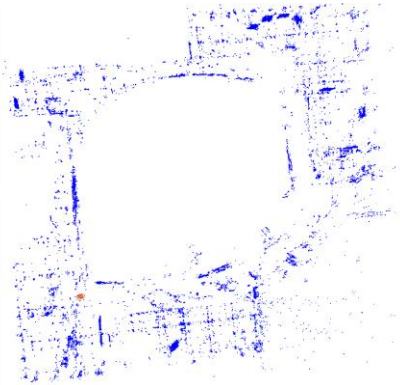
Other examples: **Face recognition for social, gesture recognition for control media app., object & post recognition for augmented reality**

Microsoft

wearable that can see!

Looxcie, Inc

who?



where?

what?

Video credits:
Matthai Philipose,

poor latency & jitter can "kill" many important mobile application

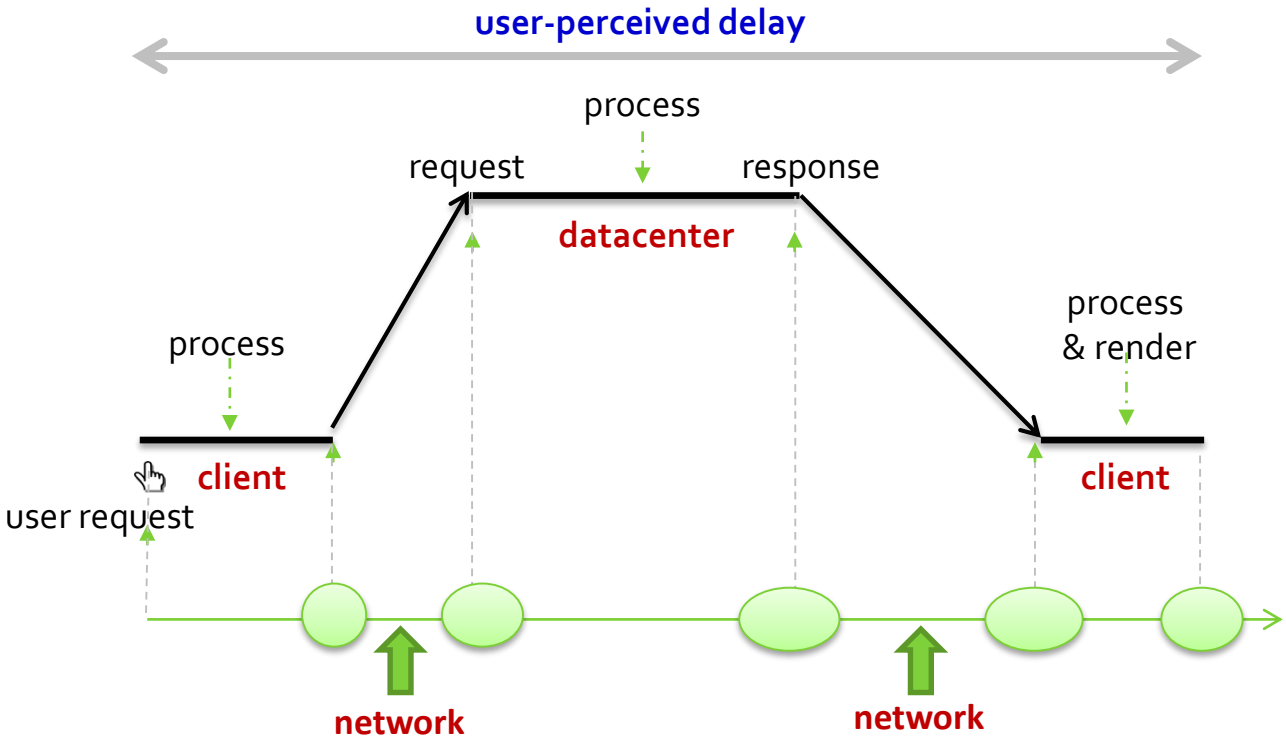
Microsoft

ground truth

- better quality Internet path enable better services
- high latency & jitter degrades services making them unusable
- poor performance impacts revenue and turns users away

components of latency

- client
- network
- datacenter



previously we focused on datacenter networking

- full bisection bandwidth networks, software load balancers, inter-datacenter SDNs

but we haven't done much with the Internet

Microsoft

no control on how packets are routed

Expt.1 : SmartPhone 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
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
10. (72.14.239.18) 22.256 ms
11. **(209.85.225.99)** 19.973 ms 21.930 ms 21.656 ms

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

Expt. 2: SmartPhone via cellular : 25 hop

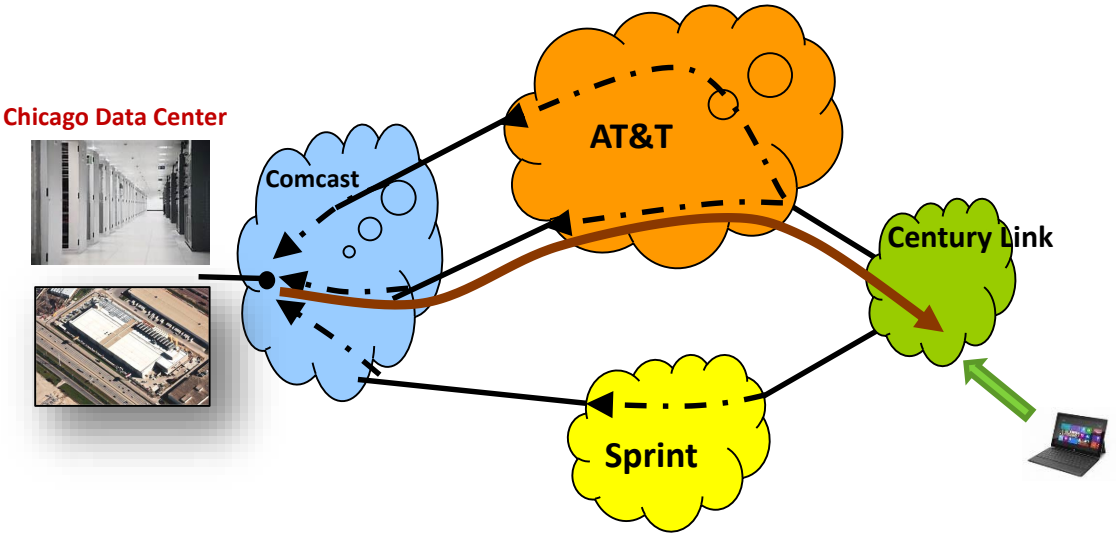
3G -> 209.85.225.99

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
14. (4.69.136.149) 889.946 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
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
21. (4.79.208.18) 669.405 ms 629.574 ms
22. (209.85.240.158) 769.538 ms
23. (209.85.241.22) 769.665 ms
24. (209.85.241.29) 589.710 ms
25. (209.85.225.99) 716.000 ms

Internet is complex

a network of networks of networks

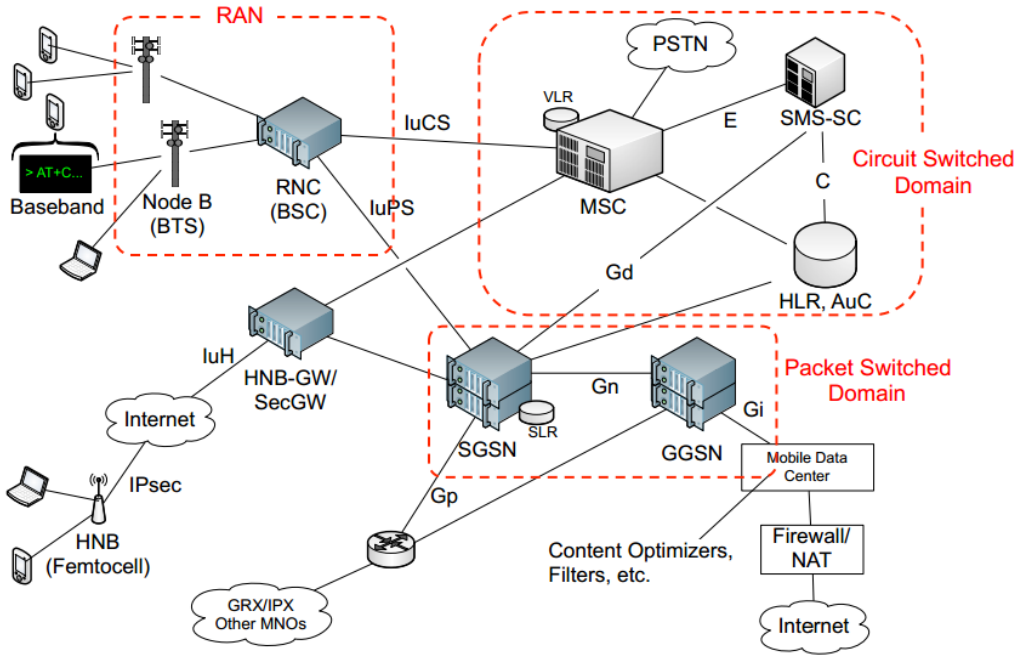
a collection of many autonomous systems (AS) managed by many ISPs with complex peering relationships

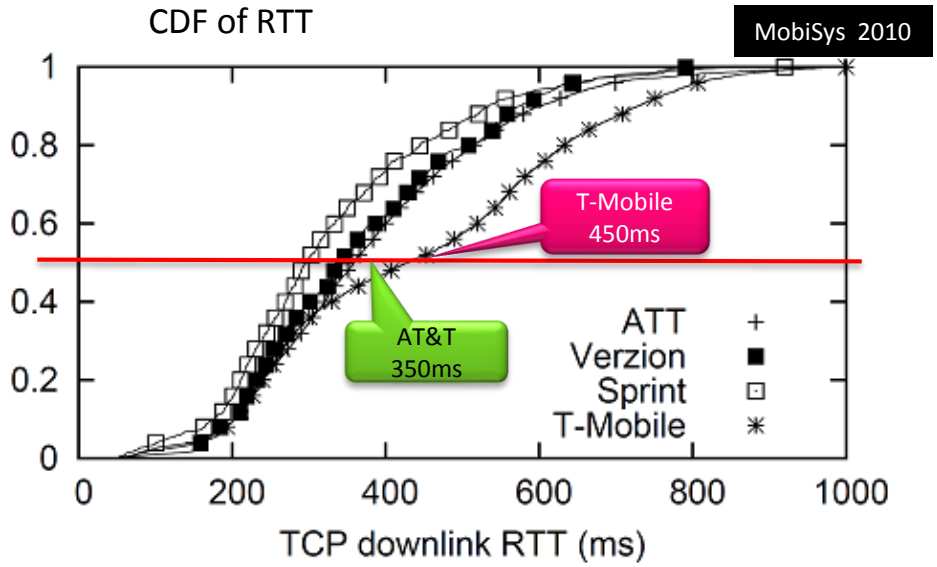


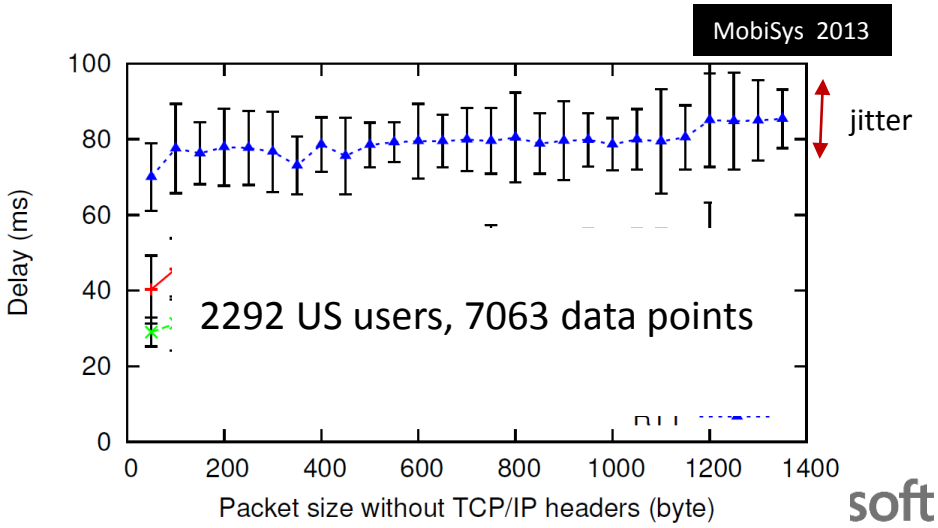
as of March 6, 2013 (source: PEER 1)

- 22,961 AS numbers (AS numbers uniquely identify networks on the Internet, e.g. 8075 for Microsoft)
- 50,519 peering connections

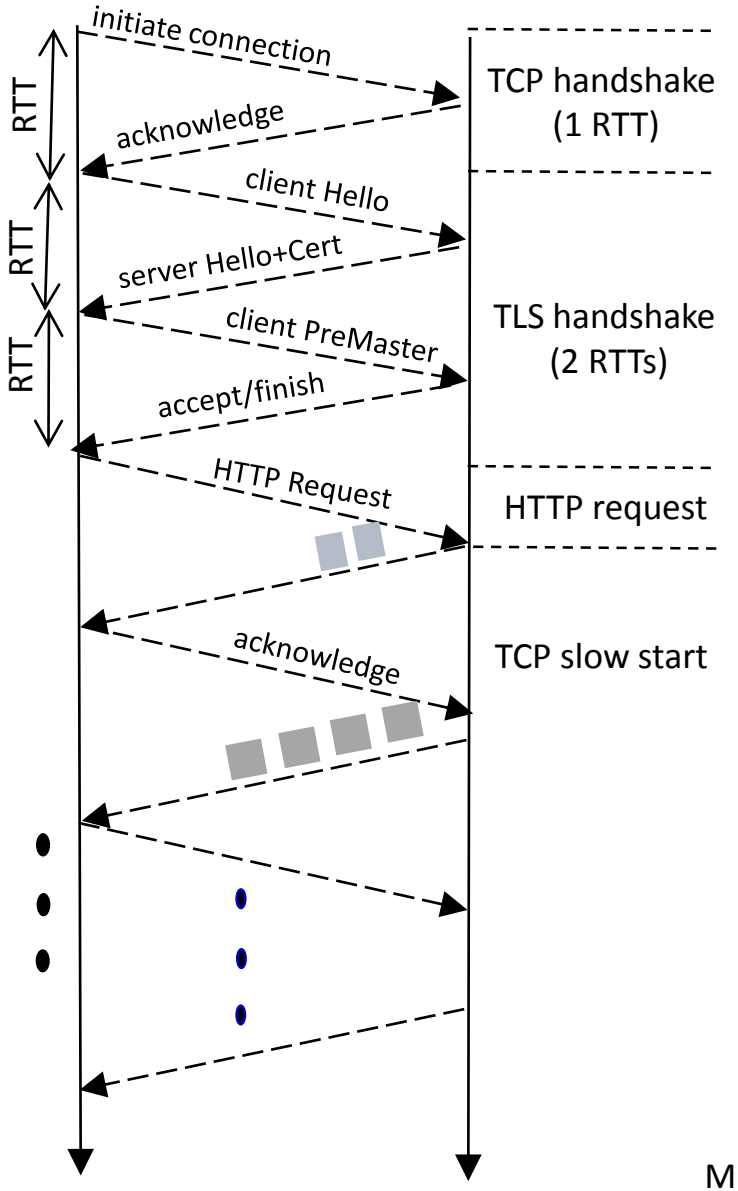
add to this the complexity of cell networks

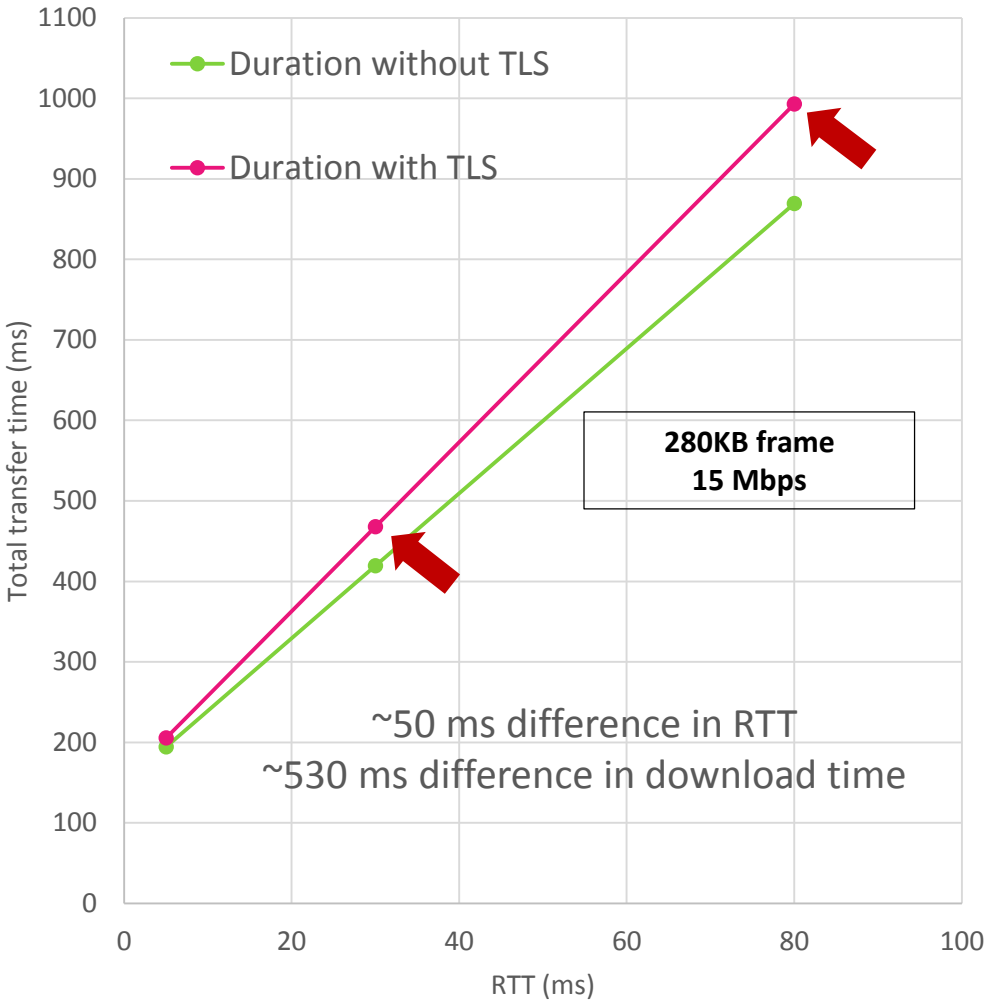






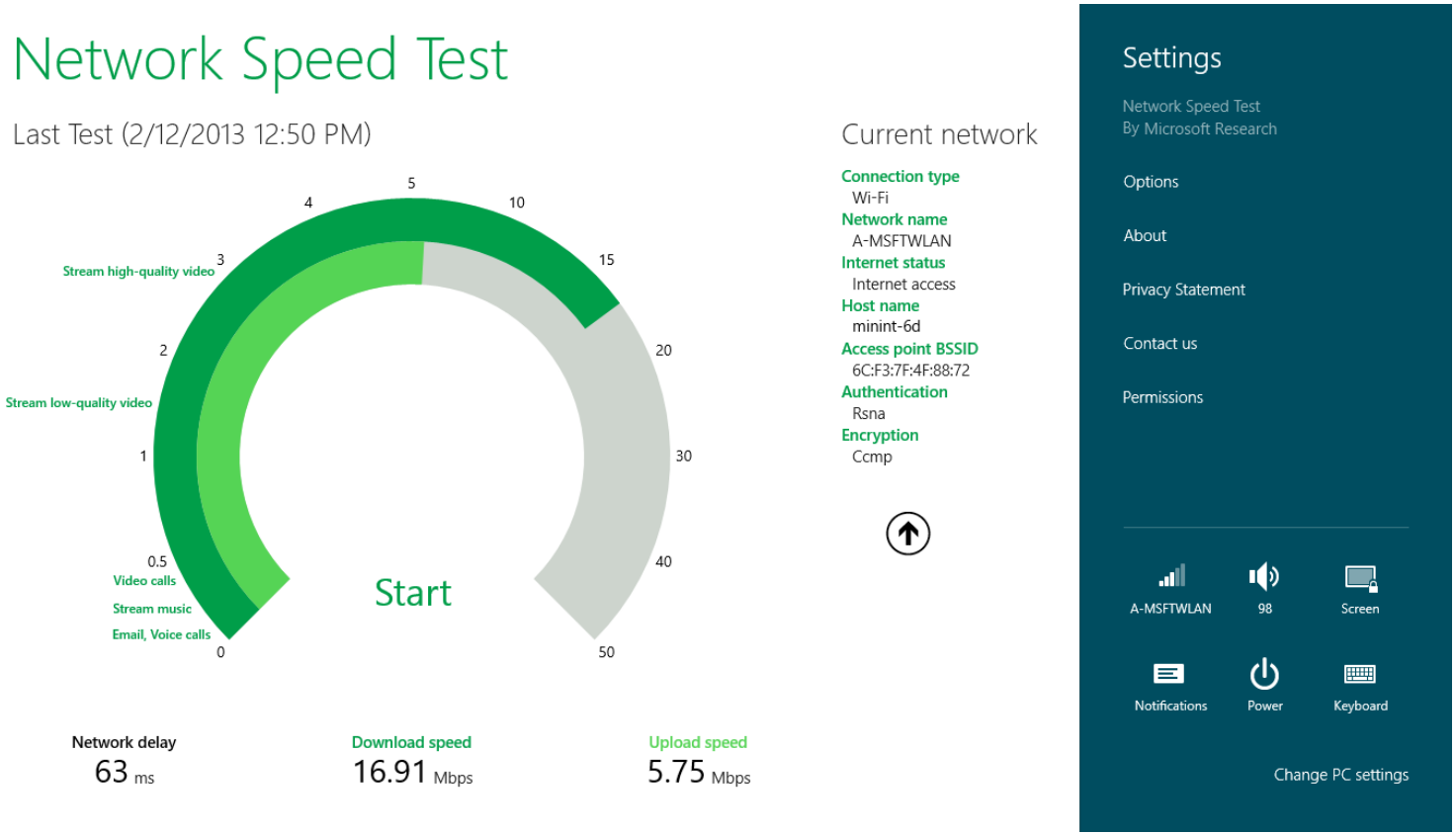
TCP & TLS make things worse





Windows MarketPlace

try it out – download Network Speed Test



Available on Windows Phone and Windows 8

Microsoft

reducing latency

possible solution

get the packets under our control as soon as possible

how?

- bring the cloud closer to the end-user
  - ✓ build lots of DCs around the world & place them in strategic locations

and large companies (Microsoft, Google, Amazon etc.) are doing just that

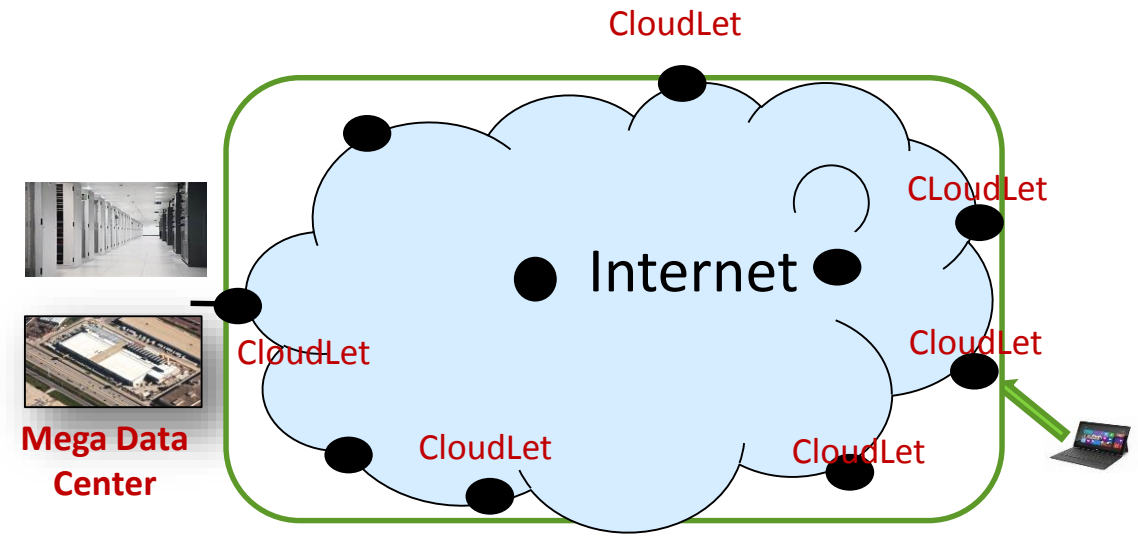
Microsoft

is building datacenters enough?

no, it's capital intensive and expensive to operate

smarter approach:

- build an extensive infrastructure of Cloudlets (4 to 40 servers with several TBs of storage, $30K-$200K/each) & place them in strategic locations



+

tunnel with strong SLAs from selected CloudLet to DCs

Microsoft

Cloudlets

definition -
a resource rich computing infrastructure with high-speed Internet connectivity to the cloud.

the mobile device uses this infrastructure to augment its capabilities and to enable applications that were previously not possible

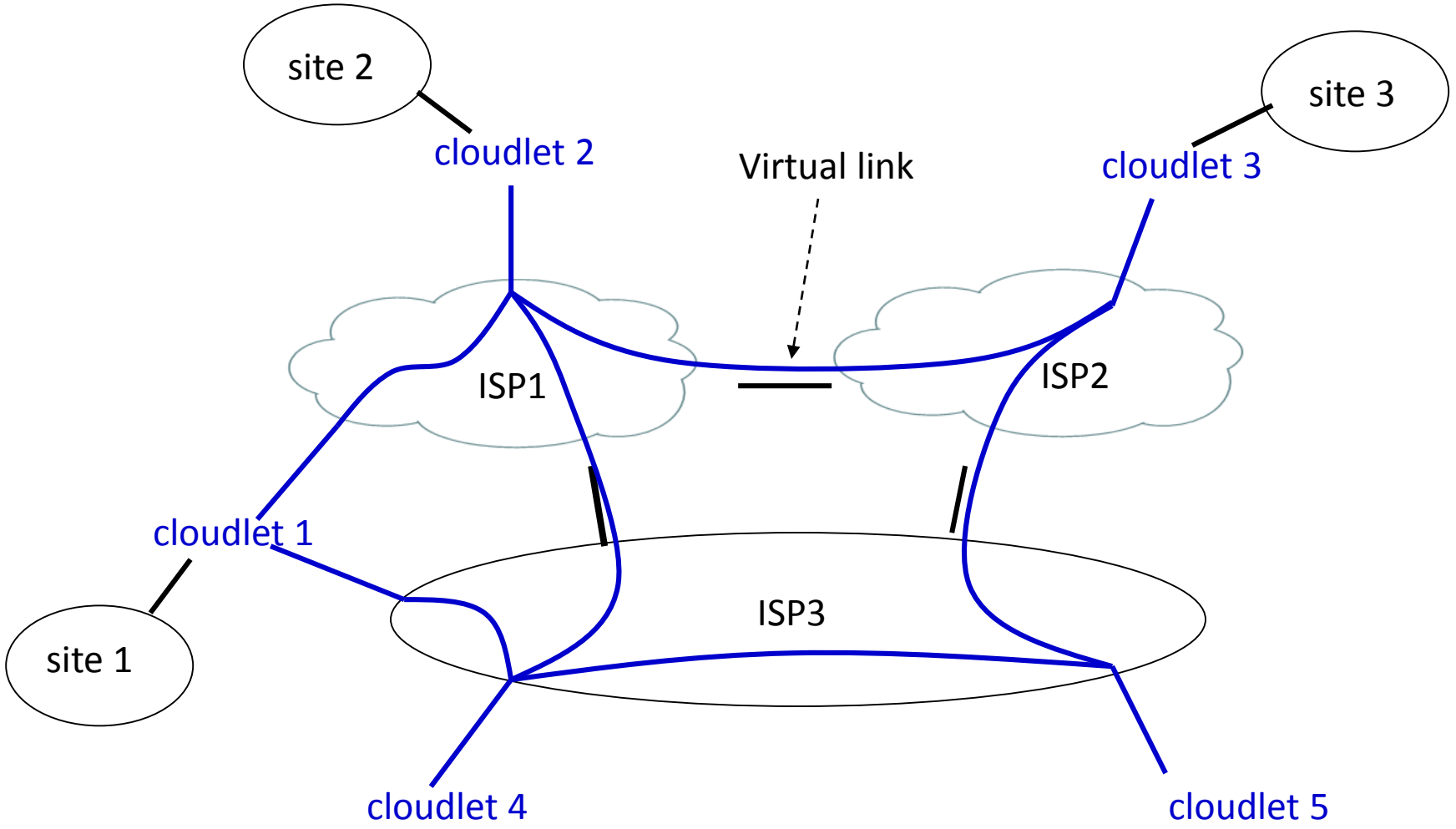
what are cloudlets good for?
site acceleration (classic)

- content caching
  - XBox, YouTube, NetFlix videos, Windows Updates,…
- split TCP connections
  - from Bing data, on avg. can reduce latencies by ~30 msec
    - predictive search query responses improved ~25-35% based on random sampling before and after deploying edge serves in a couple of US cities
- Overlay routing & path diversity

Akamai
Limelight
CloudFront
Level 3
EdgeCast
Rackspace
:
:

cloudlets are “classic” CDNs nodes, that can improve the performance of search engines, office productivity tools, video and audio conferencing & future cloud services

overlay routing leads to better paths



cloudlets exchange measurement information and choose routes.

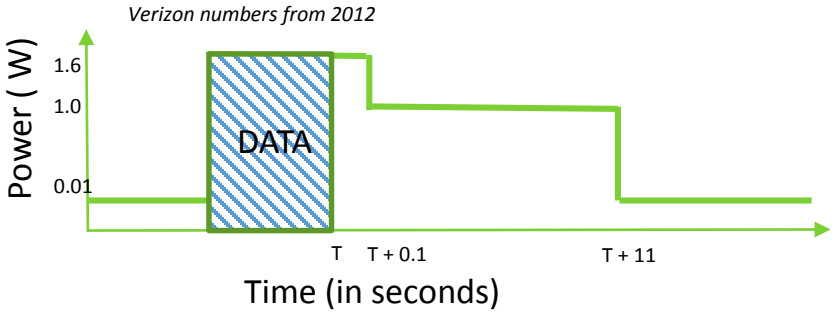
Microsoft

but cloudlets can help with battery life as well

fast dormancy

network latencies negatively impact battery life:

- LTE consumes > 1.5W when active
- LTE chip active for ~10 secs of extra tail time (1W power)



….but how did we get here

Microsoft

a bit of context/history…4 years ago

4 years ago …

The New York Times

Customers Angered as iPhones Overload AT&T

By JENNA WORTHAM
Published: September 2, 2009

The New York Times

DIGITAL DOMAIN

AT&T Takes the Blame, Even for the iPhone's Faults

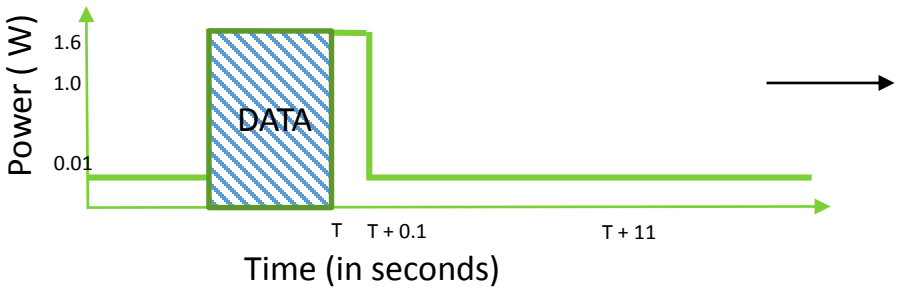
By RANDALL STROSS
Published: December 12, 2009

Report: AT&T Reputation Tarnished by iPhone Flaws

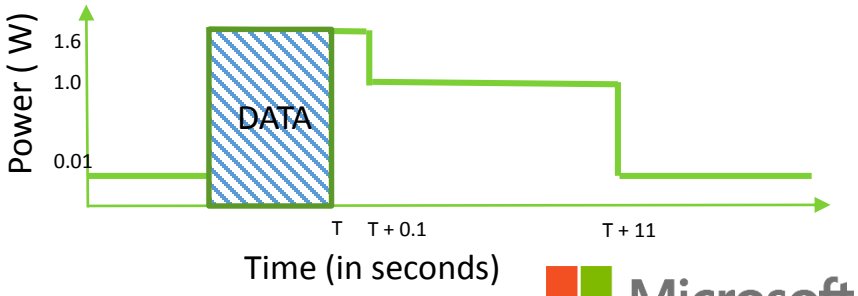
By Tony Bradley, PCWorld Dec 14, 2009 2:01 PM

original design: bring radio to low power state immediately

Verizon numbers from 2012



Mobile Operator (MO) requirement: keep LTE chip **active for ~10 sec**. of extra tail time (to reduce the signaling load)



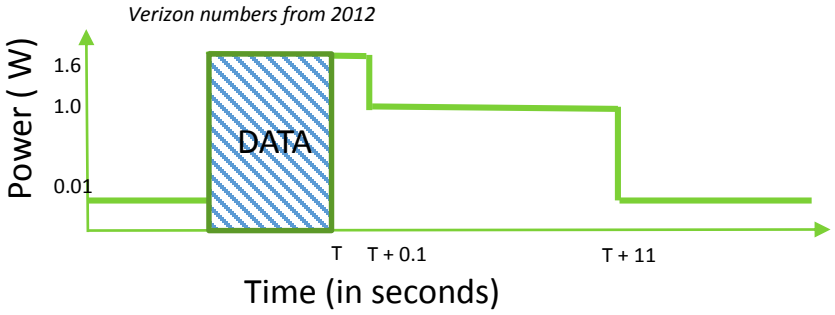
Microsoft

cloudlets can help with battery life as well

fast dormancy

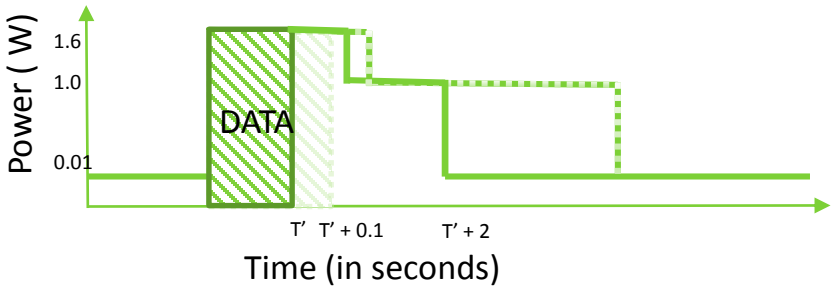
network latencies negatively impact battery life:

- LTE consumes > 1.5W when active
- LTE chip active for ~10 secs of extra tail time (1W power)



with Cloudlets:

- faster transfers => less time in highest power state
- UE can aggressively enter lowest power state



Energy savings / transfer: 1.6W*speedup + 1W*9sec = 10.6J (assuming speedup of 1 second)

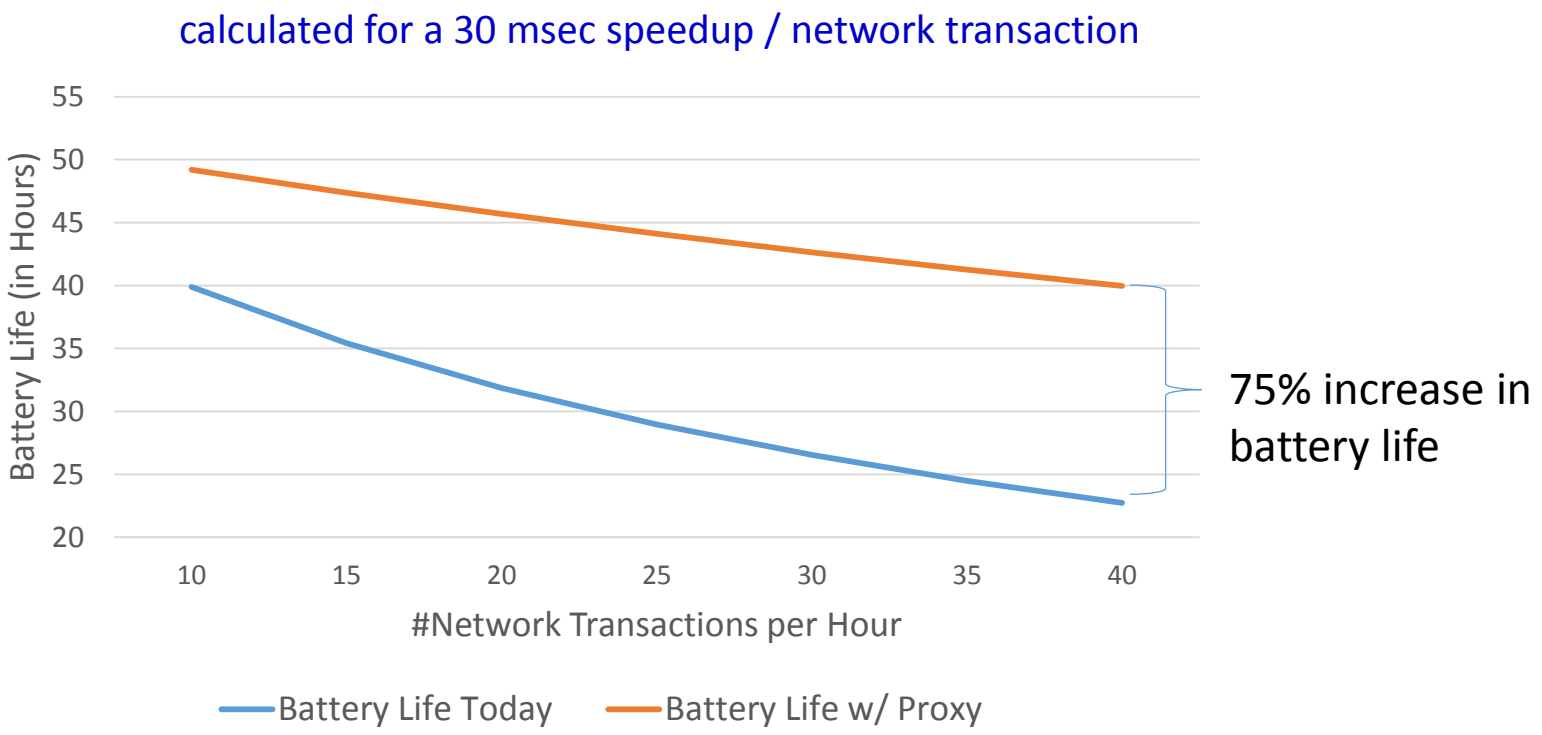
for 20 network transfers per hour (notifications, email, etc.), with 1 sec speedup, energy savings per 24 hr. = 6624 J

→ Saving of **26%** in a 1500 mAH cell phone battery*

* Samsung Standard LI-ION battery with rating of 1500mAh/3.7Vdc

Microsoft

especially good for mobile
battery life improvement



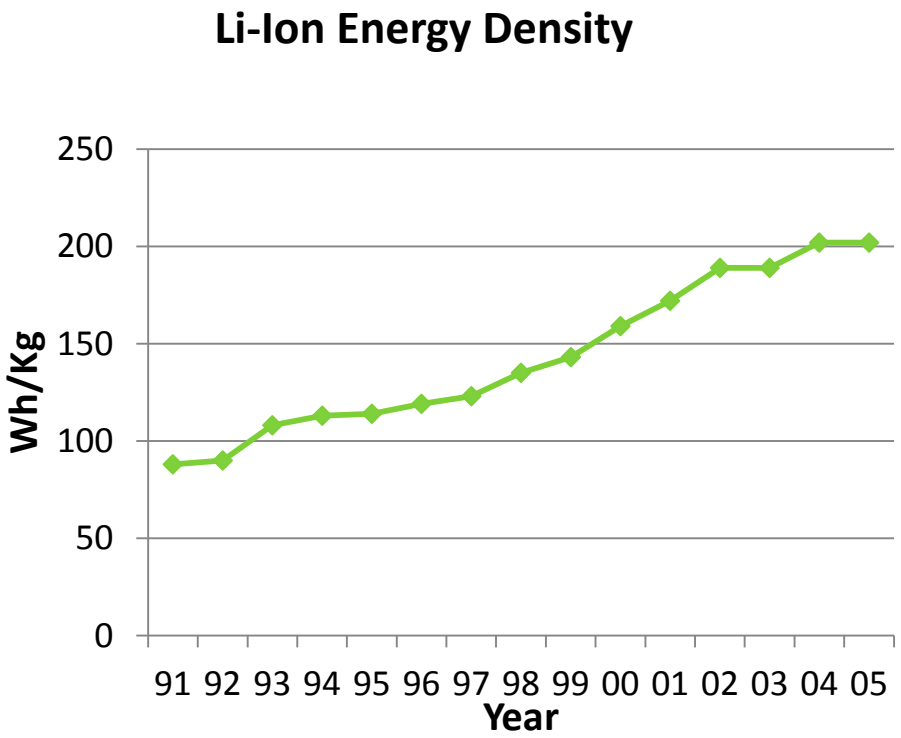
these types of saving occur across the board for all battery types and all types of mobile devices

* Samsung Standard LI-ION battery with rating of 1500mAh/3.7Vdc

Microsoft

compare to battery improvement trends

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

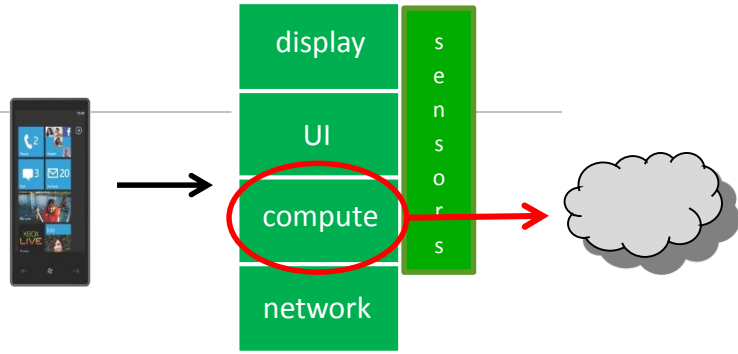
- Fast charging = lower capacity
- Slow charging = higher capacity

contrast with

CPU performance improvement during same period: 246x

Microsoft

cloudlets are great for computation offload



assertion

- remote execution reduces energy consumption and improves performance

open issues

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

I just want to write game logic on the server – I don't want to be concerned with scaling, DBs, figuring out how many servers I need, etc.

-- Game Developer-Magazine (Survey of Mobile & Social Technology, May 2012 Issue)

Microsoft

programming choices

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

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

MAUI, CloneCloud, Odessa all have a **profiler** & a **solver**

Microsoft

MAUI: 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

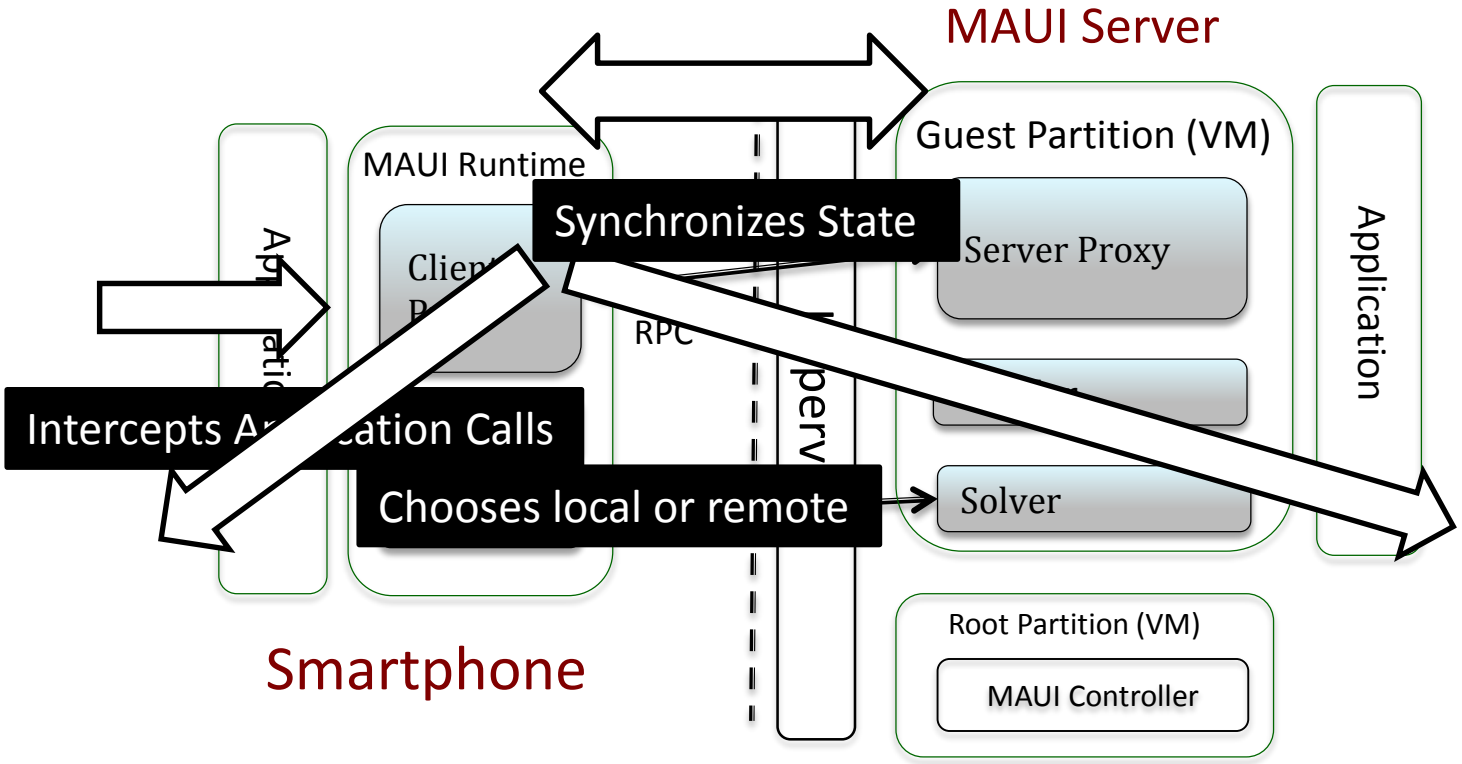
- MAUI runtime: partitions (splits) the program at run-time
  - Can optimize for energy-savings, or performance

why not use a static client/server split?

- developers need to revisit application structure as devices change
- 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

programming model: MAUI (Cont'd)

Application Partitioning



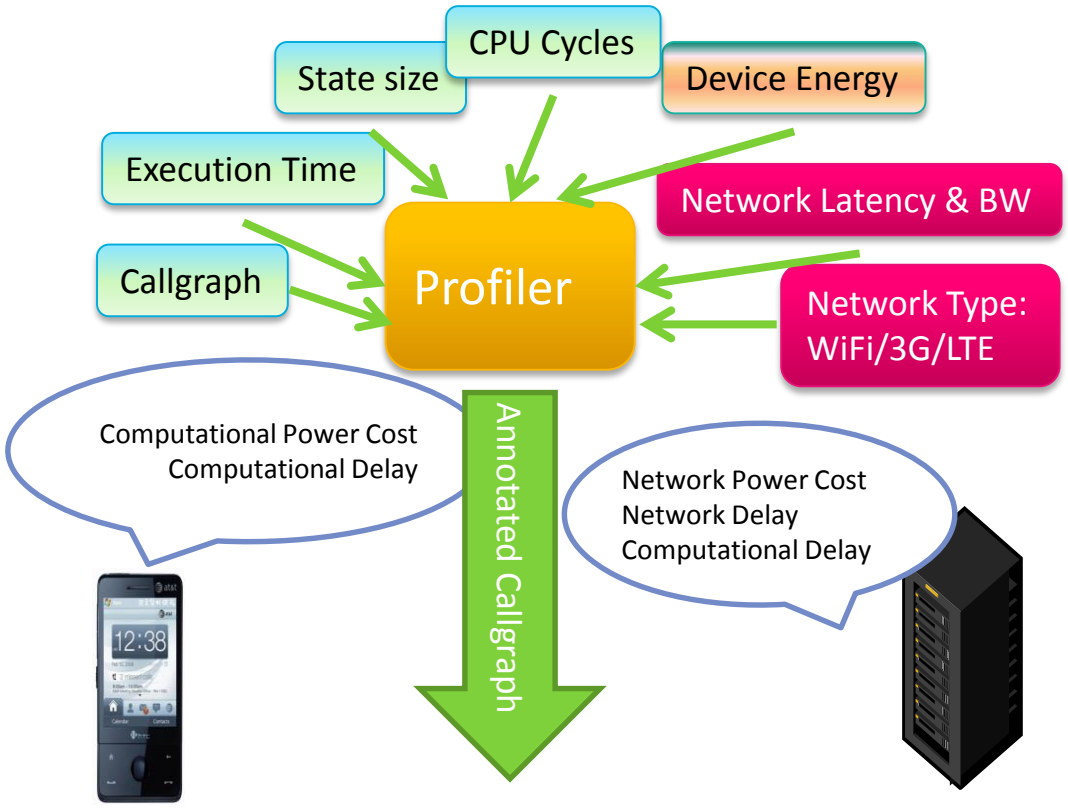
client/server split, can be extended to multiple tiers

Microsoft

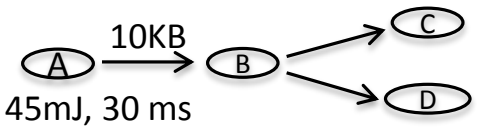
profiler and decision engine

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

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

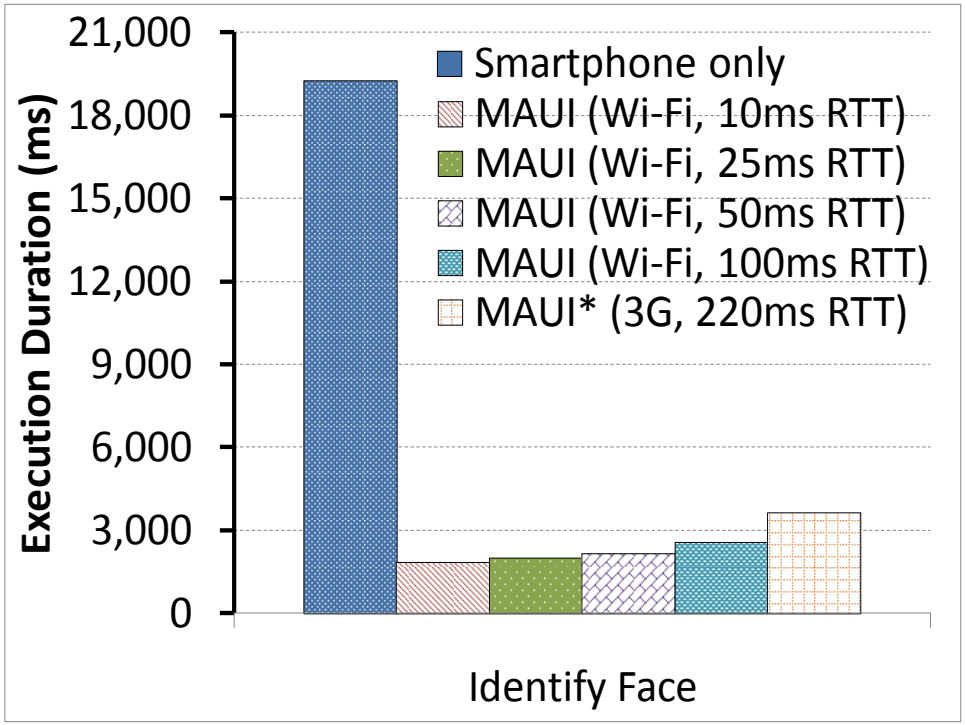
$I_v <= R_v$ for all $v \epsilon V$

Microsoft

performance and energy benefits

Performance Benefits:

Memory Assistant Face recognizer:



Face recognition becomes “interactive” w/ offload

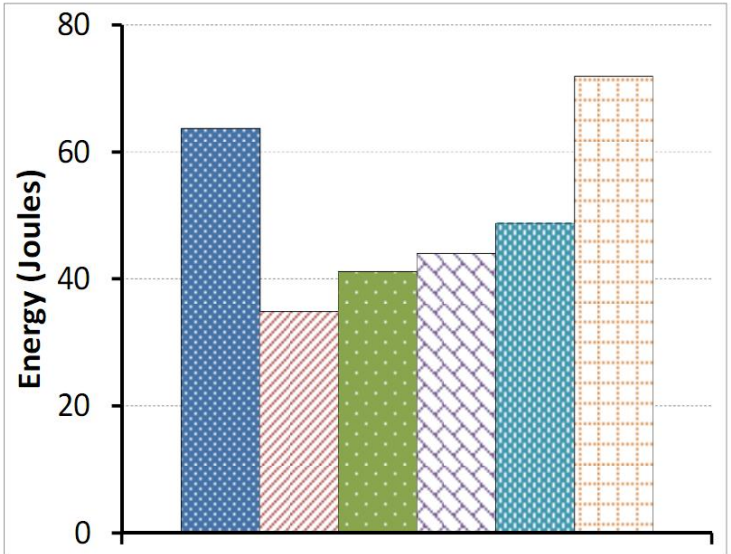
Energy Benefits:

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

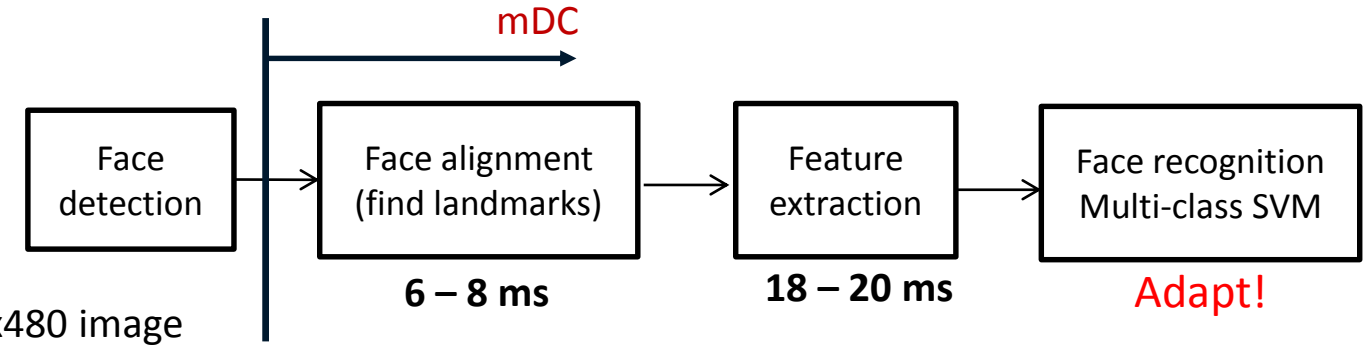


cloud offloading

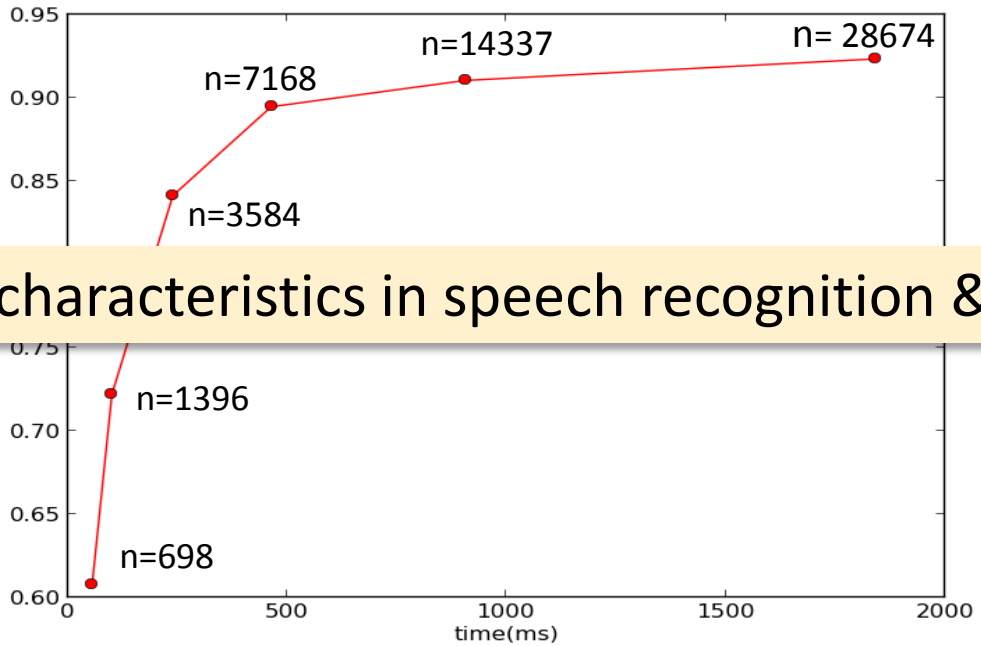
augmented reality

the lower the latency, the better the results

new services: object recognition



For a 640x480 image
Client: 766ms
Server: 138ms



Face prediction Time

similar characteristics in speech recognition & search

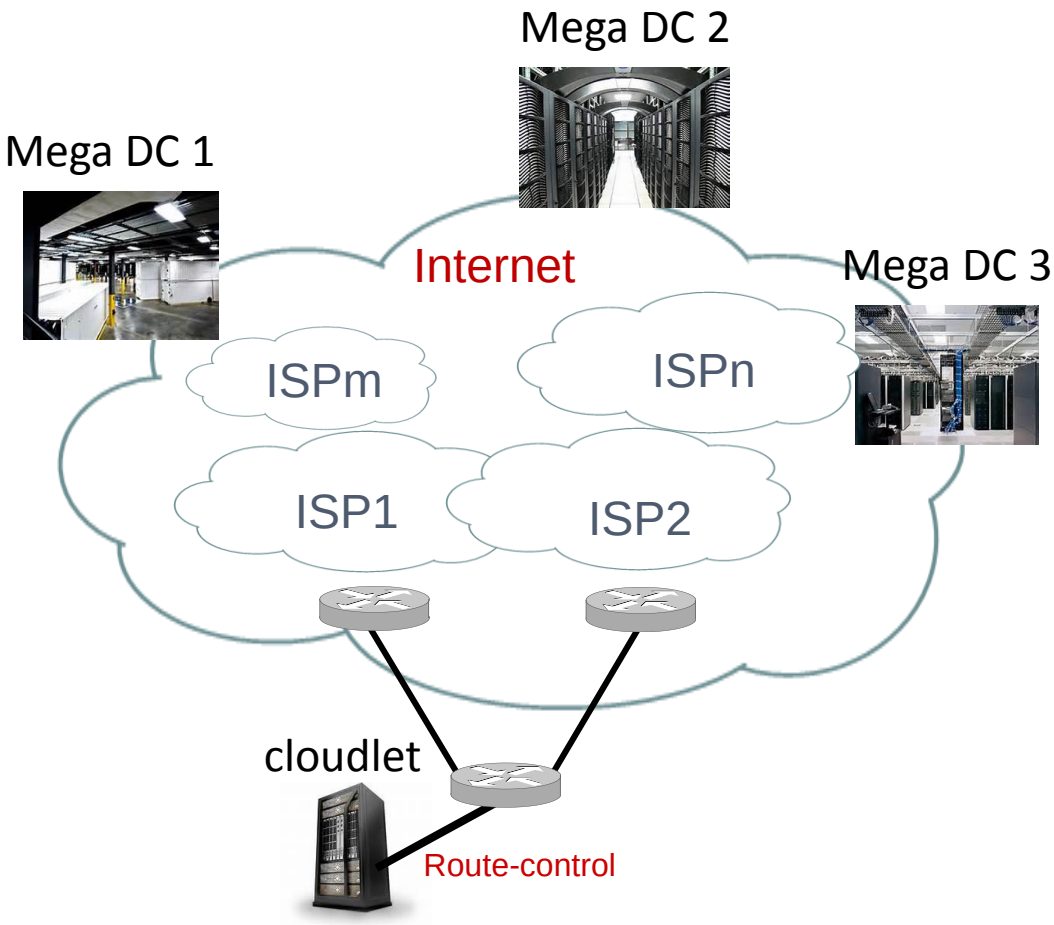
Microsoft

cloudlets are also good for resilient connectivity

- tunnels with SLAs between cloudlets and DCs enable better mobile experience & better performance
- overlay routing via cooperating cloudlets in different routing domains do better by re-routing through peer nodes SOSP 2001
- path diversity via multi-homed cloudlets improves Internet performance USENIX 2004 SIGCOMM 2003

Microsoft

multi-homing leads to better paths



Microsoft

cloudlets can reduce dependency on cellular networks

offload to Wi-Fi aggressively } already doing this

compress aggressively } e.g. WP London + TPG optimizations

NEW

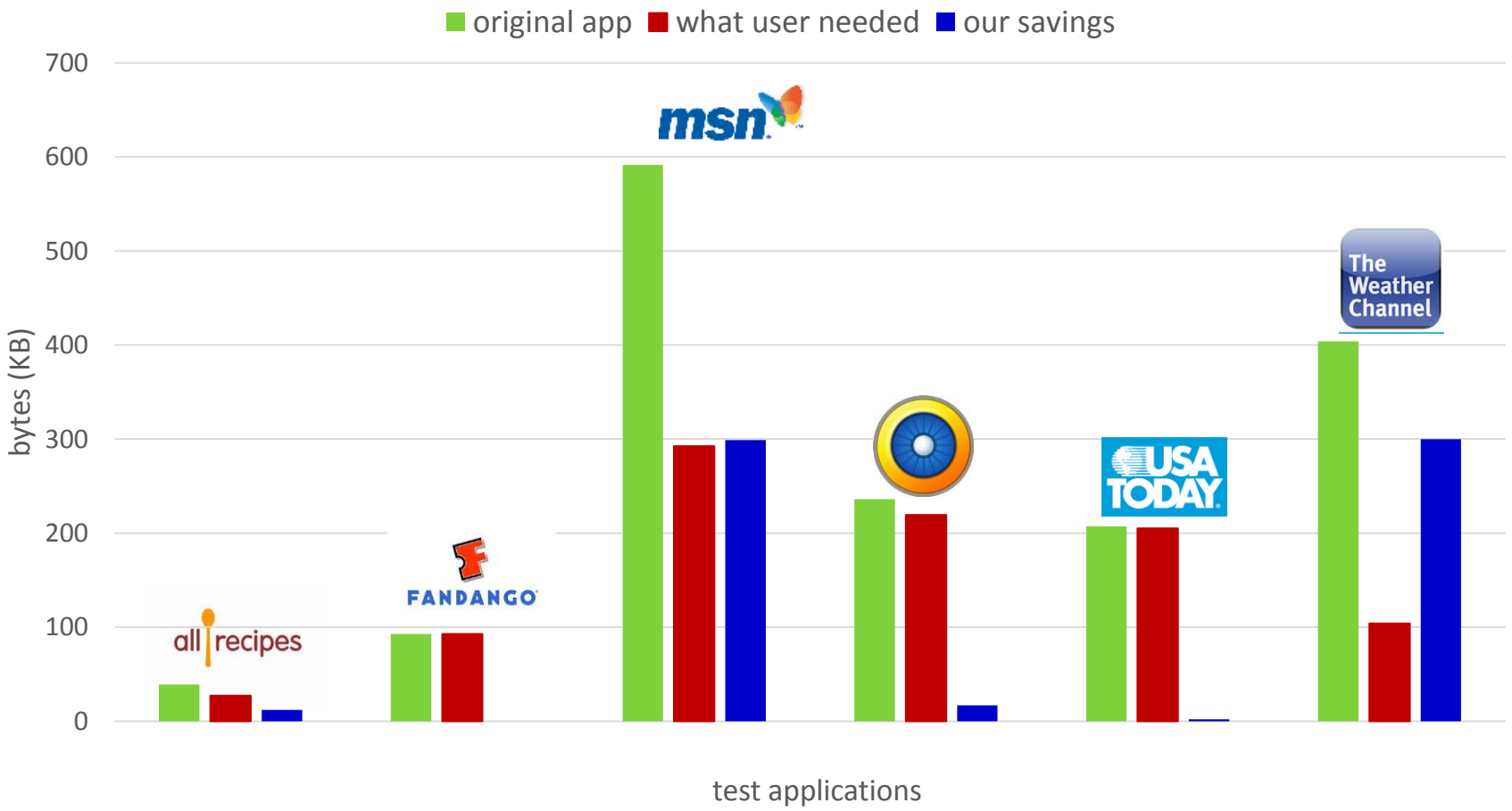
procrastinate instead of prefetch

- many network apps. fetch data whether or not it is consumed
- idea: mDC fetches the data but holds on to it until user explicitly needs it
  - ✓ save cellular bandwidth without the latency penalty

procrastinate & save

few results on bandwidth saving

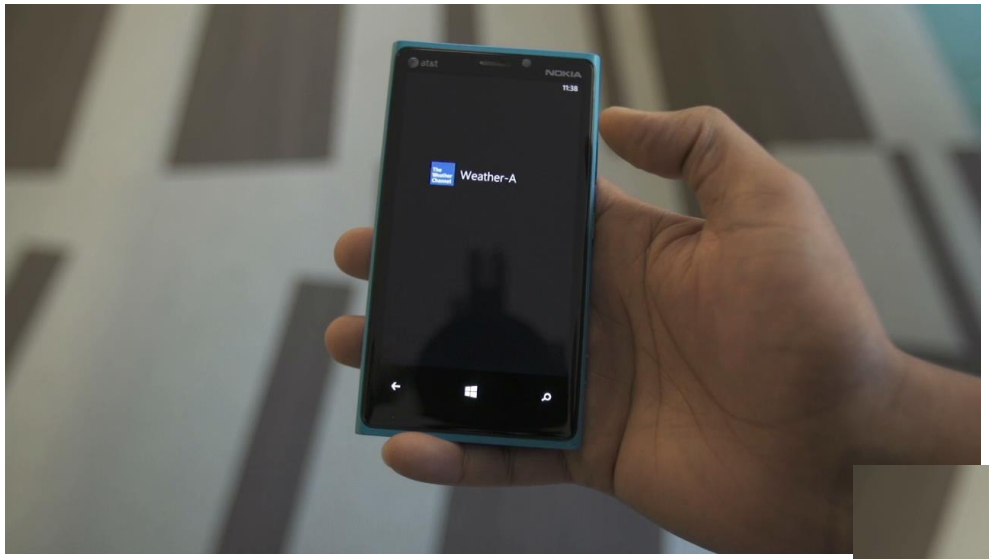
the system automatically decides what is not needed by the end-user



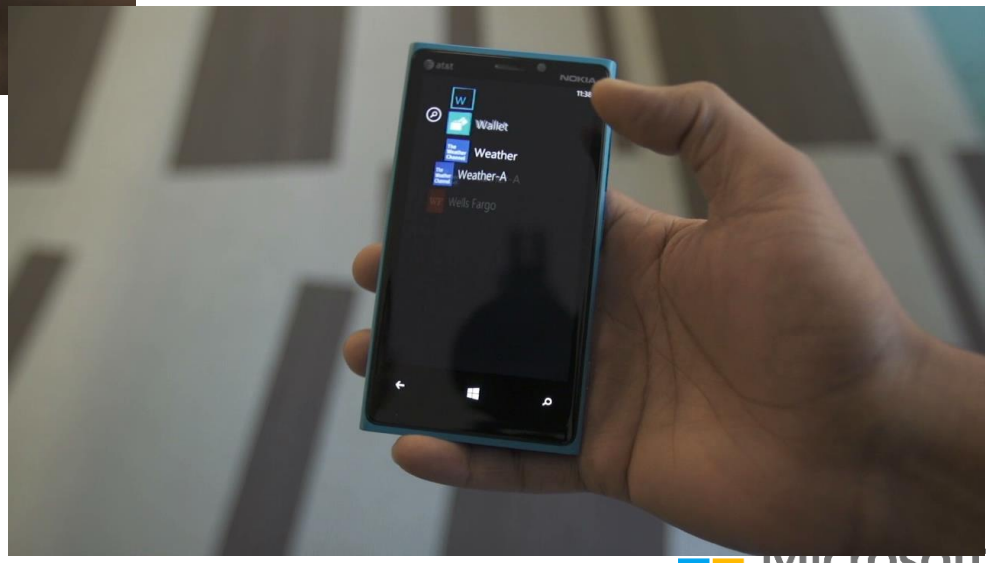
micro datacenter - benefits

reducing dependency on cellular networks (with procrastination)

get data only when needed (without mDC)



get data only when needed (with mDC)



cloudlet benefits -
app streaming & game streaming

run any ecosystem's apps on our devices by streaming them from the cloud

- circumvent client-side compatibility complexities
  - apps are hosted just like office 365

NEW mDCs reduce

- latency -- keeping users engaged
- jitter & packet loss – reduce user frustrating in highly interactive sessions
- backbone bandwidth so both MOs and we pay less to other ISPs

note: standard proxy + split TCP insufficient for interactive traffic

other important cloudlet services

- VM virtualization / isolation for multi-tenancy
  - partnerships can be formed, different parties can pay for deployed infra-structure
- service and Internet monitoring
  - faster detection & localization of network problems – can lead to CSS cost reduction
- improved IP2GEO localization (for targeted ads.)

computing at the edges was the hot topic of discussion in a recent NSF workshop on "*Future Directions in Wireless Networking 2013*"

Microsoft

summarizing benefits of cloudlets

latency reduction

- ✓ caching - serve static content immediately)
- ✓ SSL termination / split TCP
- edge to DC protocol enhancements

bandwidth saving

- ✓ compression
- procrastination

service & internet monitoring

reliable connectivity

- overlay networking
- path diversity

battery saving

- client proxying
- procrastination
- computation offloads

app streaming

- reduce the app gap
- lower device cost

Revenues

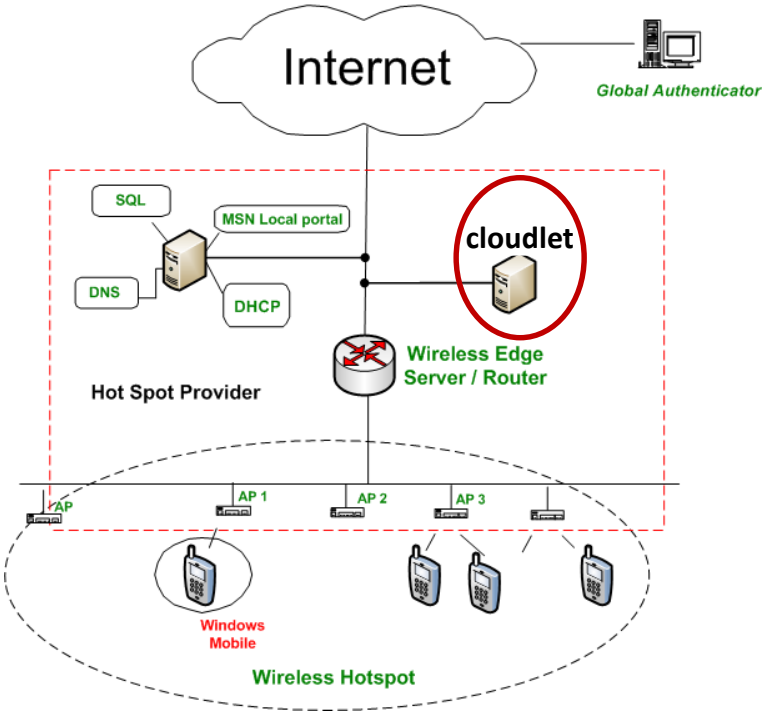
- locally relevant advertisements
- multi-tenancy

denial-of-service protection, new services & reduction of load on DCs

deployment

Wi-Fi is an excellent choice for cloudlets

Enterprise & publics spaces



Wi-Fi LAN vendors

ABIresearch
technology market intelligence

ARUBA networks

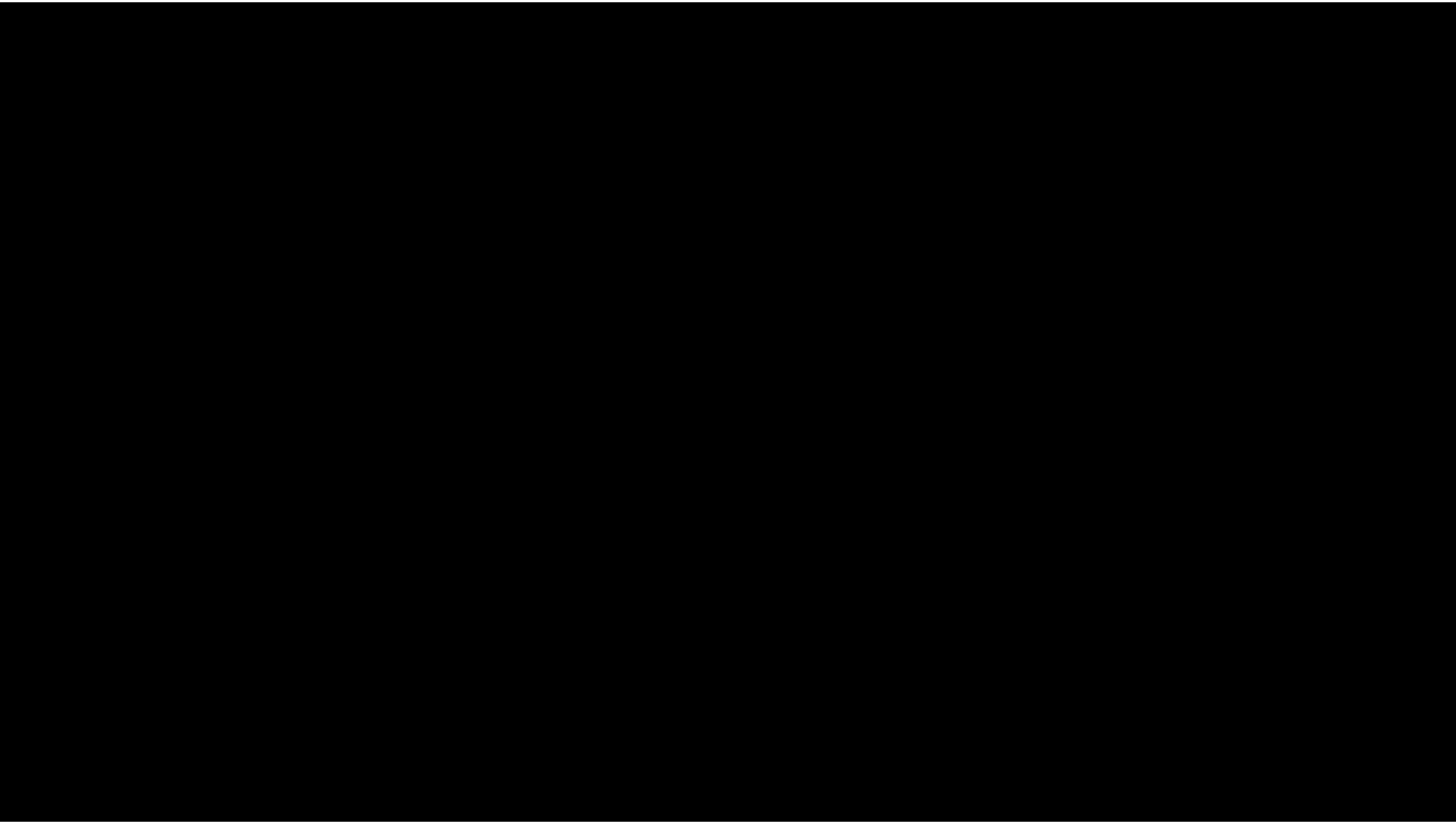
CISCO

Ruckus
Simply Better Connections.

BROADCOM.

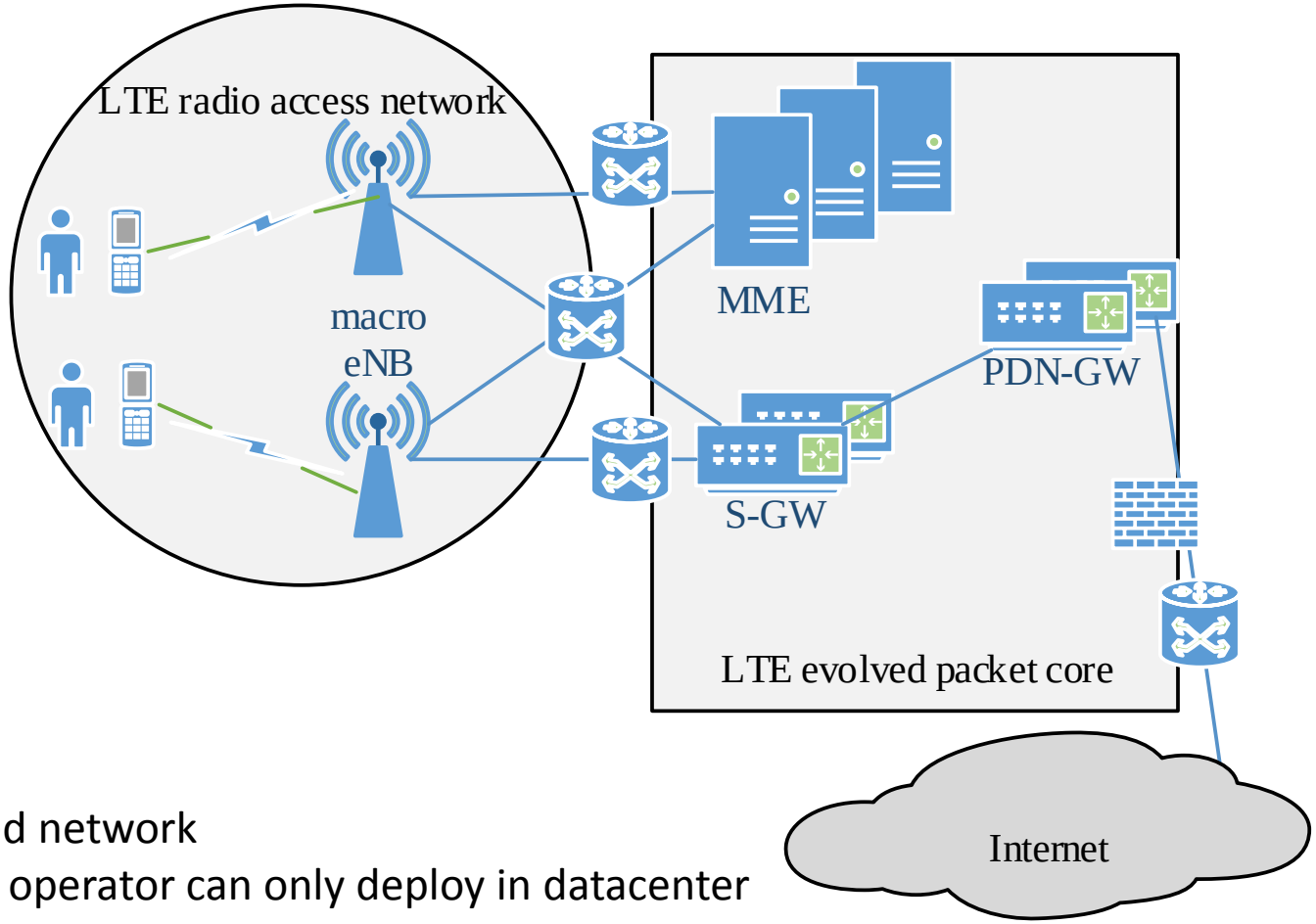
Microsoft

especially good for fast action cloud gaming



Microsoft

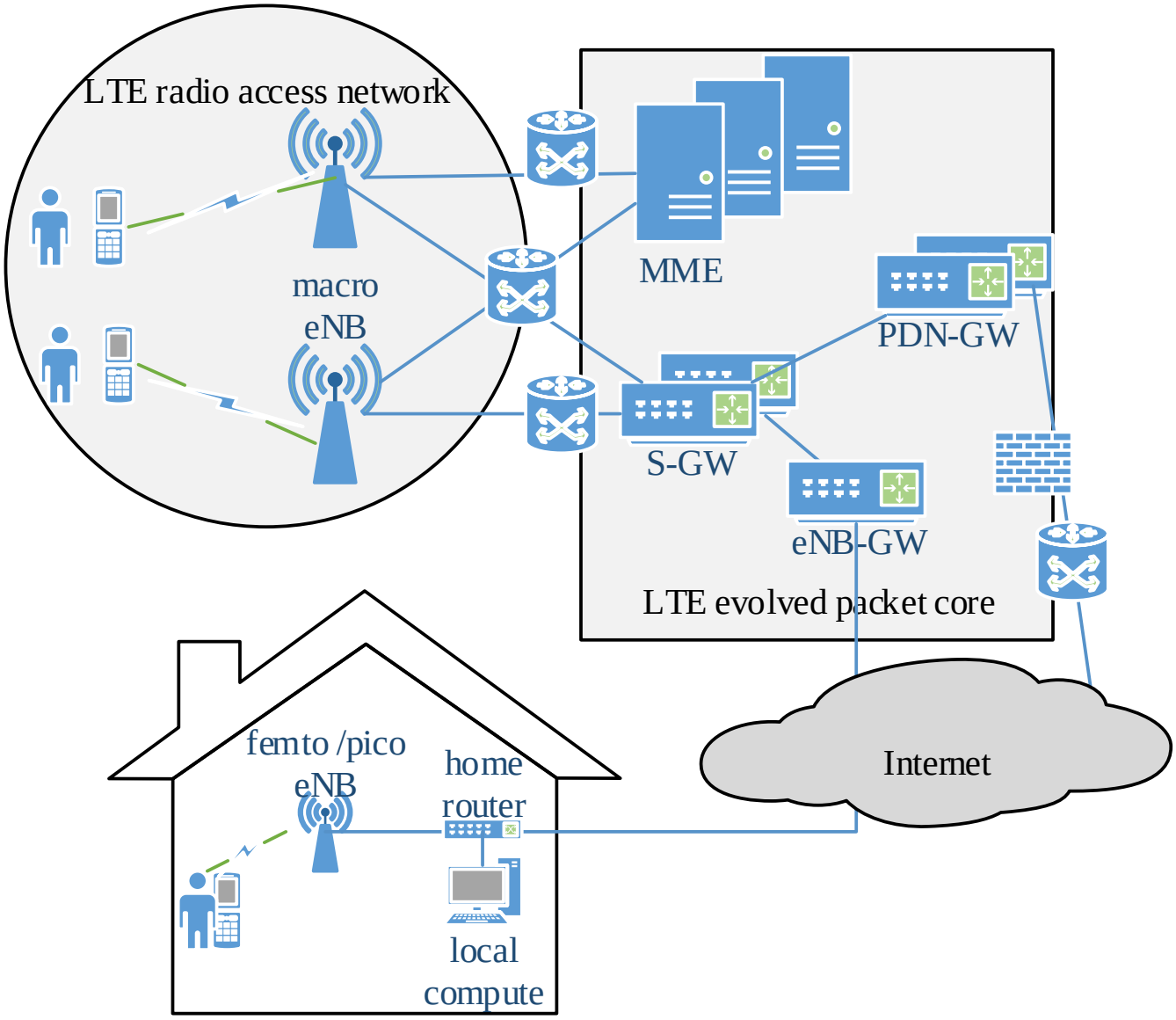
cellular networks … not so much



closed network
even operator can only deploy in datacenter
LTE – 70ms, 12mbps DL, 5 mbps UL

Microsoft

small cells



Microsoft

why even consider small cells?

similar footprint, size, cost to Wi-Fi AP

billing, authentication, SMS, voice

- just works

licensed frequencies

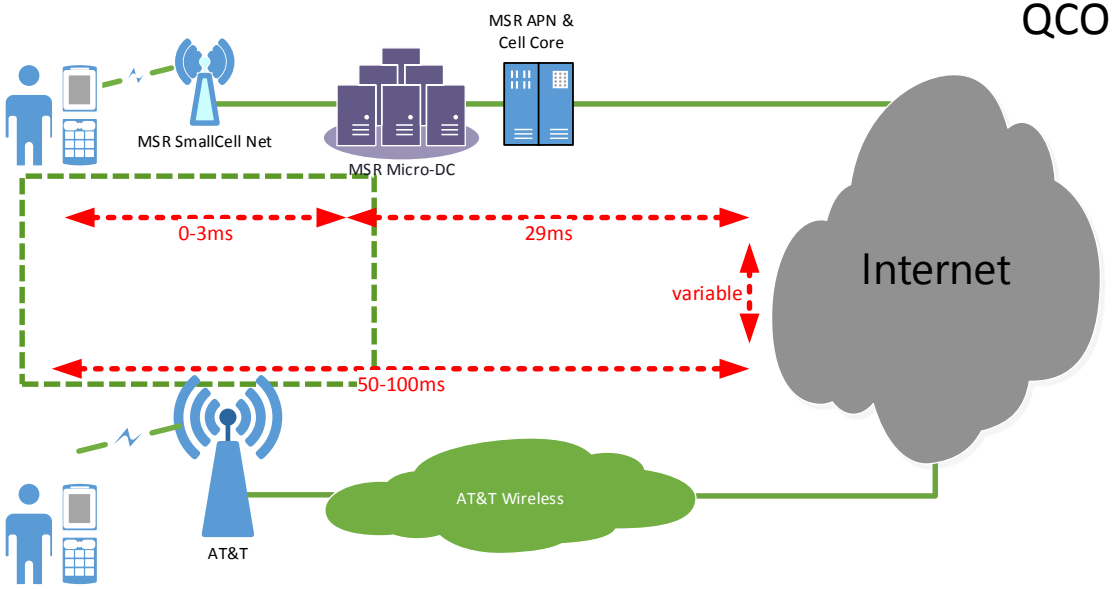
- interference only from other cells & devices
- SON for frequency reuse, power control
- handoff works

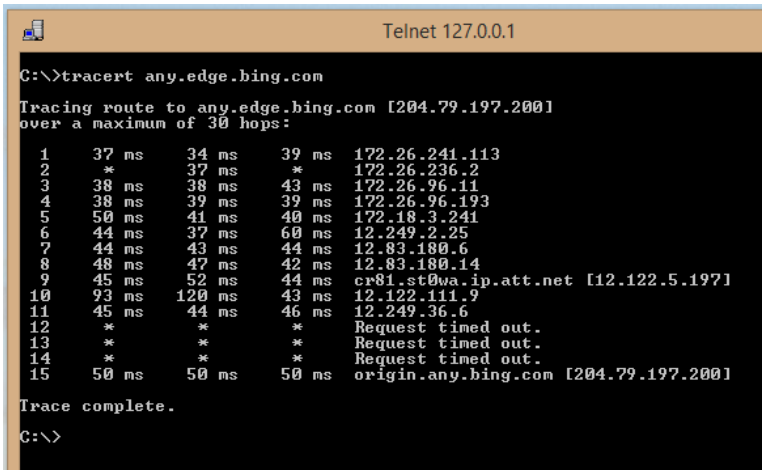
Microsoft

our experience with small cells

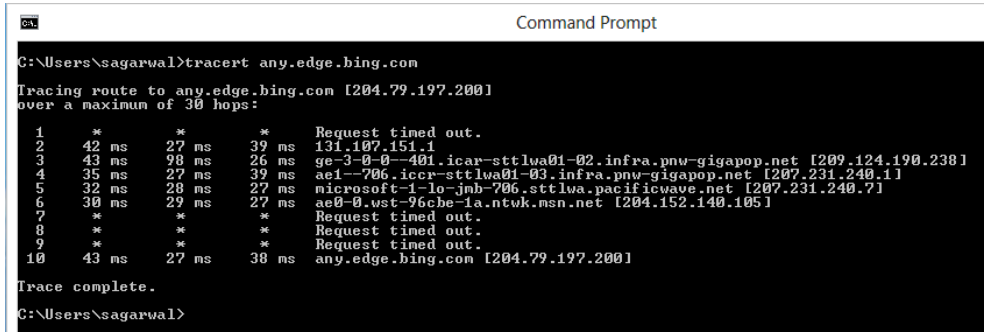
everything is faster

QCOM's Small Cell





tracert from AT&T LTE to *any.edge.bing.com* *(15 hops)*



tracert from SC to *any.edge.bing.com* *(10 hops)*

Microsoft

LTE performance

metric	median	25th %	75th %
DL throughput	12.6 mbps	7.6 mbps	19.7 mbps
UL throughput	5.5 mbps	1.9 mbps	11.2 mbps
RTT	71 ms	50 ms	98 ms

small cell performance (Huawei)

metric	value
DL throughput	∼110 mbps
UL throughput	∼10 mbps
RTT	∼11 ms

small cell growth

Informa, Feb 2013, "Small cell market status"

- In 2012 the no. of SC deployed overtook the total no. of macrocells
- Sprint reported 1+ million units during Oct 2012
- AT&T estimate also 1 million
- 9 of top 10 mobile operator groups (by revenue) are offering femtocell services, incl. AT&T, China Mobile, France Telecom/Orange, Telefonica, T-Mobile/ Deutsche Telekom and Vodafone

PR Newswire
A UBM plc company

[October 30, 2013]

Small Cells & Femtocells Market Worth $5.98 billion by 2019

Fig. 8: Small-cell ecosystem, 3Q12

Source: Small Cell Forum

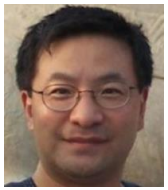
other's are thinking about "cloudlets" as well

Increasing Mobile Operators' Value Proposition With Edge Computing

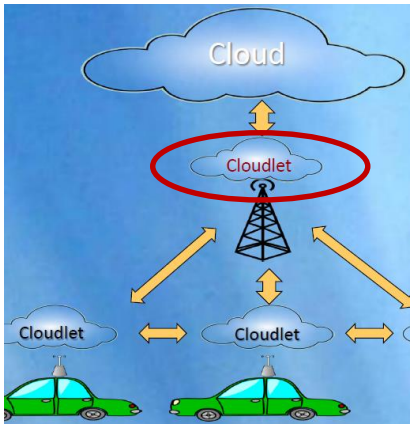
Turn bit pipes into smart pipes with an Intel® architecture-based server embedded into a Nokia Siemens Networks* base station

"local cloud are essential for backbone and core network scalability"

Dr. Geng Wu, Chief Scientist, Intel (Wireless World Research Forum, Vancouver, BC, Oct. 22, 2013)



5G with **Undelay Networks** and **Local Cloud**



"cloudlets for reducing latency, security and reliability"

- Dr. David Soldani, VP Huawei Research Centers (IEEE ICC, June 12, 2013)

Microsoft

others are thinking about cloudlets as well

NETWORKWORLD

News | Blogs | Newsletters | Videos | Events | INSIDER | More

Security | LANs & WANs | UC / VoIP | Cloud | Infrastructure Mgmt | Wireless | Software | Data Center | SMB

Broadband | Collaboration | Equipment | Mobile | Networks | Security | Storage | White Papers | Webcasts | Tests

News

Nokia Siemens to merge cloud, base-station computing to boost performance

The company's Liquid Applications platform will use computing power in the cloud and in base stations, based on conditions

By Stephen Lawson, IDG News Service
February 24, 2013 04:06 PM ET

Add a comment | Print | Share | +1 | Like

IDG News Service - Nokia Siemens Networks will expand the role
with a new platform that will store and deliver some application da
into information about subscribers and traffic to improve the proce

The company announced the system, called Liquid Applications,
on the eve of Mobile World Congress. Liquid Applications can imp
experience but cutting delays as well as delivering more relevant

IBM

United States [change]

News room | Search

Home | Solutions | Services | Products | Support & downloads | My IBM | Welcome [IBM Sign in] [Register]

News room
- News releases
- Press kits
- Image gallery
- Biographies
- Background
- News room feeds
- Global news rooms
- News room search
- Media contacts

News room > News releases >

IBM and Nokia Siemens Networks Announce World's First Mobile Edge Computing Platform

News release | Contact(s) information | Related XML feeds

Barcelona, Spain - 25 Feb 2013: Mobile World Congress: Nokia Siemens Networks and IBM (NYSE: IBM) announced today a collaboration to deliver the world's first mobile edge computing platform that can run applications directly within a mobile base station. This new platform allows mobile operators to create a truly unique mobile experience, relieve the ever increasing strain on network infrastructure and bring completely new solutions to market.

The new platform can accelerate the delivery of media-rich services by delivering content directly from the base station, ensuring enhanced quality of experience for consumers in the face of ever increasing data traffic growth. The platform also enables a new generation of

Contact us
- Contact a media relations representative
- Site feedback

Share
- Facebook
- Twitter

Microsoft

moving MOs towards offering edge services ("OCDNs")

Why a Cloudlet Beats the Cloud for Mobile Apps

Posted on December 13, 2009 by lewisshepherd

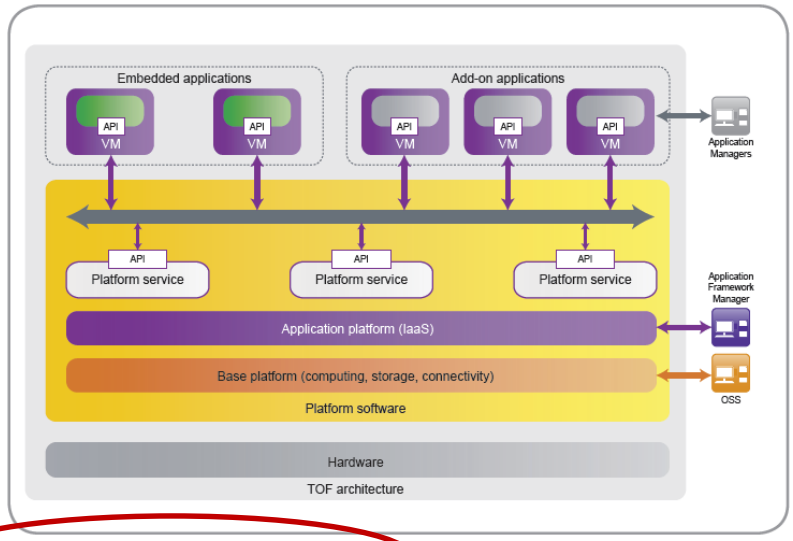


Figure 4: Base station application architecture

Liquid Net

Nokia Siemens Networks

there is plenty of research literature (incl. MSR's) that shows edge computing significantly enhances mobile experience

Microsoft

cloudlets in MO networks

A classic reason:

- caching,…

The CDN case

top CDN companies have moved in this direction:

GIGAOM

Akamai Going Mobile With Velocitude Buy

by Ryan Lawler JUN. 10, 2010 - 8:17 AM PST

ZDNet

Akamai buys Cotendo for $268 million, eyes mobile cloud

Summary: *Given that Akamai is a leading content delivery network (CDN) it's clear that Cotendo's focus on mobile services fits in well.*

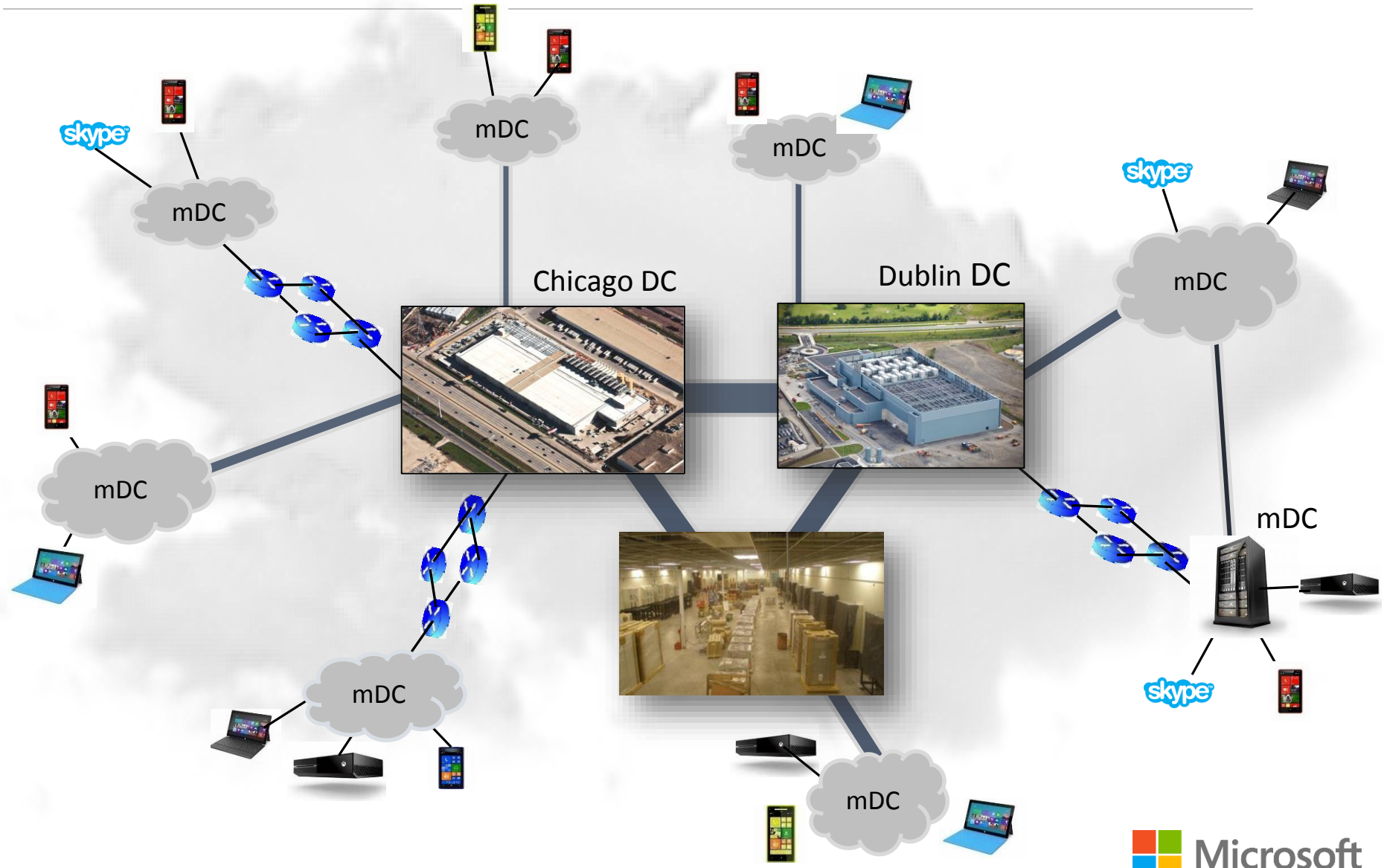
By Larry Dignan for Between the Lines | December 22, 2011 -- 04:23 GMT (20:23 PST)

Microsoft

summarizing

- cloudlets = classic CDNs + multi-tenant edge computing + overlay networking
- cloudlets = 1 to 40 servers with high speed, high-bandwidth connectivity well connected to mega DCs (clouds)
- mobile computing can get a serious boost from distribution of cloudlets on the internet

food for thought: the future of cloud computing is the disaggregated cloud (with lots of open questions)



Thanks!