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 (desktops, servers)
- weight, power, size constraints
- CPU, memory, display, keyboard

finite energy supply
- actions, decisions must be deferred
- huge hardware & wireless networking improvements since but deep essentials still the same. Will the same slide will be true in 2020?

communication uncertainty
- bandwidth / latency variation
- intermittent connectivity
- may cost real money, require service agreements
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

Li-Ion Energy Density

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

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 the programmer
  – Programmer builds app as standalone phone app
  – Programmer adds .NET attributes to indicate “remoteable” methods / classes

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

Decision Engine:
Partition A Running App

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

Example – Maximize:
\[
\sum_{v \in V} (I_v \times E_v) - \sum_{(u,v) \in E} (|I_u - I_v| \times C_{u,v})
\]
energy saved cost of offload

Such that:
\[
\sum_{v \in V} (I_v \times T_v) + \sum_{(u,v) \in E} (|I_u - I_v| \times B_{u,v}) \leq \text{Lat.}
\]
execution time time to offload

and

\[
I_v \leq R_v \text{ for all } v \in V
\]

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

Performance Benefits:
Memory Assistant Face recognizer:

Energy Benefits:
Interactive arcade game w/physics engine:

Face recognition becomes “interactive” w/ offload

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]
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.
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

Wouldn’t it be cool to print from a phone...
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

<table>
<thead>
<tr>
<th>Game Type</th>
<th>Latency Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>First-person, Racing</td>
<td>≈ 100 ms</td>
</tr>
<tr>
<td>Sports, Role-playing</td>
<td>≈ 500 ms</td>
</tr>
<tr>
<td>Real-time Strategy</td>
<td>≈ 1000 ms</td>
</tr>
</tbody>
</table>
the matchmaking problem

End-to-end Latency Threshold

Connection Latency

Clients

Match to satisfy total delay bounds
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
- Deployment
- Processing
- Model Evaluation
- Cloud Services

- OEM firmware
- OTA Updates

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

- Cloud Services
user adaptation

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

\[
p_{\text{adapt}}(w_i | w_{i-1}w_{i-2}) = \lambda_1 p_{\text{user}}(w_i | ...) + \lambda_2 p_{\text{recent}}(w_i | ...) + \lambda_3 p_{\text{static}}(w_i | ...)
\]

\[
\sum_i \lambda_i = 1
\]

Learned on the client

Shipped by OEM
Determining \( \lambda \)

**Optimal Static Weight \( \lambda \)**

23% relative gain over Mango

This is just a baseline!

Static LM only

User LM only
next word prediction + fixed candidates

DEMO

Microsoft Confidential
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
Curved to match thumb

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
- Activity Classification
- SensD
- Image Stitching
- Snakes & Dragons
Microsoft Research
Project Hawaii

Sample Press Articles

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.

Microsoft Research delivers cloud development kit for Windows Phone

Microsoft's new Hawaiian codenames are all about mobile

Developers: Windows Phone 7 + Cloud Services SDK By MSR

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

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

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

All services are integrated with Azure MarketPlace
phones that see

who?

where?

what?

red apple

Video credits:
Matthai Philipose, MCRC
Intel Labs
what about connectivity?
bandwidth scarcity: demand continues to go up


Industry Forecasts of Mobile Data Traffic

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)

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 Marketplace
65 Reviews, averages review rating of 4.75/5 stars
### 3G --> 209.85.225.99

1. ***(172.26.248.2) 414.197 ms 698.485 ms 539.776 ms**
2. **(172.16.0.66) 539.781 ms 809.954 ms 689.547 ms**
3. **(12.122.138.21) 669.690 ms 529.758 ms 699.965 ms**
4. **(12.88.242.189) 589.857 ms 1129.848 ms 709.784 ms**
5. **(12.122.138.38) 589.781 ms 1009.723 ms 769.808 ms**
6. **(10.251.11.23) 689.837 ms 669.340 ms 689.739 ms**
7. **(10.251.10.2) 509.781 ms 729.746 ms 679.787 ms**
8. **(198.110.131.78) 12.715 ms 9.424 ms 9.315 ms**
9. **(192.122.183.41) 11.381 ms 6.054 ms 5.975 ms**
10. **(192.12.80.69) 7.038 ms 7.353 ms 7.026 ms**
11. **(198.108.23.12) 12.525 ms 13.027 ms 12.619 ms**
13. **(192.122.183.41) 11.381 ms 6.054 ms 5.975 ms**
14. **(192.12.80.69) 7.038 ms 7.353 ms 7.026 ms**
15. **(198.108.23.12) 12.525 ms 13.027 ms 12.619 ms**
17. **(192.122.183.41) 11.381 ms 6.054 ms 5.975 ms**
18. **(192.122.183.41) 11.381 ms 6.054 ms 5.975 ms**
19. **(192.122.183.41) 11.381 ms 6.054 ms 5.975 ms**
20. **(192.122.183.41) 11.381 ms 6.054 ms 5.975 ms**
21. **(192.122.183.41) 11.381 ms 6.054 ms 5.975 ms**
22. **(192.122.183.41) 11.381 ms 6.054 ms 5.975 ms**
23. **(192.122.183.41) 11.381 ms 6.054 ms 5.975 ms**
24. **(192.122.183.41) 11.381 ms 6.054 ms 5.975 ms**
25. **(209.85.225.99) 716.000 ms 669.435 ms 559.875 ms 499.598 ms**

---

**iPhone over 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 (209.85.241.35) 124.588 ms*
10. *(209.85.225.99) 19.973 ms 21.930 ms 21.656 ms*

---

**iPhone via 3G : 25 hop**

3G -> 209.85.225.99

1. ***(172.26.248.2) 414.197 ms 698.485 ms 539.776 ms**
2. **(172.16.0.66) 539.781 ms 809.954 ms 689.547 ms**
3. **(12.88.242.189) 589.857 ms 1129.848 ms 709.784 ms**
4. **(12.122.138.38) 589.781 ms 1009.723 ms 769.808 ms**
5. **(10.251.11.23) 689.837 ms 669.340 ms 689.739 ms**
6. **(10.251.10.2) 509.781 ms 729.746 ms 679.787 ms**
7. **(198.110.131.78) 12.715 ms 9.424 ms 9.315 ms**
8. **(192.122.183.41) 11.381 ms 6.054 ms 5.975 ms**
9. **(192.12.80.69) 7.038 ms 7.353 ms 7.026 ms**
10. **(198.108.23.12) 12.525 ms 13.027 ms 12.619 ms**

---

**tracert 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.
sample deployment scenario

augment Wi-Fi hot spots with cloudlets (in public 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

Spectrum Availability

End-user Consumption

Advanced Networks

Applications & Content

Mobile Devices

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
Thanks!
mobile computing

massive dependency / addition

Would rather give up:

- 70% Alcohol
- 63% Chocolate
- 55% Caffeine
- 54% Exercise
- 33% Sex
- 22% Toothbrush

- 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
4. All other data: Vodafone in-house research
comparing growth trends
2010 – 2015, million units

SmartPhones outsell laptops & desktops

<table>
<thead>
<tr>
<th>Year</th>
<th>Feature Phone</th>
<th>Smartphone</th>
<th>Portable PC</th>
<th>Desktop PC</th>
<th>Tablet</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>1,085</td>
<td>305</td>
<td>305</td>
<td>146</td>
<td>18</td>
</tr>
<tr>
<td>2011</td>
<td>1,085</td>
<td>472</td>
<td>201</td>
<td>147</td>
<td>50</td>
</tr>
<tr>
<td>2012</td>
<td>1,091</td>
<td>607</td>
<td>215</td>
<td>152</td>
<td>73</td>
</tr>
<tr>
<td>2013</td>
<td>1,104</td>
<td>740</td>
<td>246</td>
<td>155</td>
<td>92</td>
</tr>
<tr>
<td>2014</td>
<td>1,104</td>
<td>866</td>
<td>284</td>
<td>156</td>
<td>107</td>
</tr>
<tr>
<td>2015</td>
<td>1,096</td>
<td>982</td>
<td>332</td>
<td>158</td>
<td>121</td>
</tr>
</tbody>
</table>

’10–’15 CAGR
- Feature Phone: 0.2%
- Smartphone: 26.4%
- Portable PC: 11.3%
- Desktop PC: 1.7%
- Tablet: 46.6%

Sources: BLS, IDC, CEA, Accenture analysis
connectivity options over unlicensed frequencies

### Speed
- 10 Gbps
- 1 Gbps
- 500 Mbps
- 100 Mbps
- 10 Mbps
- 5 Mbps
- 1 Mbps

### Range
- 0
- 10
- 50
- 100
- 1000
- 10,000

### Today’s World

#### 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, ...

IEEE 802.11 {b, g, a, n}
Wireless USB
UWB
WhiteFi
IEEE 802.11af, IEEE 802.22, ECMA TC48-TG1

New World

60 GHz
- WirelessHD, WiGig, IEEE 802.11ad, IEEE 802.15.3c

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

IEEE 802.11p

IEEE 802.15.1
Bluetooth
IEEE 802.15.4
Zigbee
IEEE 802.15.4f
RFID
the world’s first urban white space network

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
FCC Officials Visit Microsoft To Examine Experimental Network

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)

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

High loss rate & low throughput severely limits the scope of cloud services
putting the cloud in the palm of your hands

Victor Bahl
5.23.2012
from vision to strategy

going 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)