Towards Smart Digital Home Storage

Young Jin Nam

CSE Department, University of Minnesota
School of Computer & IT, Daegu University
Talk Outline

1. Data Growth @ Home
2. Analysis of Our Digital Home (requirements)
3. Analysis of Existing Home Storage
4. Summary & Future Work
Digital Data Growth @ Worldwide

- "Digital Universe" from 2009 to 2020
  - data: 44 times, $800K \text{ PB}('09) \rightarrow 35,200K \text{ PB}('20)$
  - # files: 67 times, storage capacity: 30 times
  - staff & investment: 1.4 times

(source: iView: The Digital Universe Decade – Are You Ready?, IDC 2010.5)
Good News is ...

• 75% of digital data are copies!
  – 25% are unique!
  – high chance to reduce the storage demand/supply gap!
  – but, regulations for multiple copies(reliability/availability)

• Data de-duplication with some challenges
  – mainly applied in 2\textsuperscript{nd}-tier storage (archive, backup..)
  – should work with primary storage(performance concern)
  – even with (public & private) cloud storage
Needed are ...(1/2)

• **Intelligent search tool**
  – mostly unstructured data (images, audio...)
  – how to add structure to unstructured data
  – how to find the information we need when we need it?

• **New storage & information management scheme**
  – what information we need to keep & how to keep it?
  – classify data by importance, know when to delete/need

*(source: iView: The Digital Universe Decade – Are You Ready?, IDC 2010.5)*
Needed are ...(2/2)

- **More compliance tools**
  - compliance with government/industry regulations
  - 2009, $46 billion (keeping records/transactions/privacy)

- **Better security**
  - information to be secured is growing 2x faster than data growth

(source: iView: The Digital Universe Decade – Are You Ready?, IDC 2010.5)
Digital Data Growth @ Home

• ~12 TB @ household in 2014
  – home entertainment, backup, home video...

Coexistence of HDD & Flash

- HDD (high resolution contents, backup, library...)
- Flash as mobile storage (lower resolution contents)

HDD @ Home

- Mostly in set-top-boxes & multimedia devices

- Decline in 2007 vs. 2006—particularly in mobile CE market—short term or long term trend?
- In 2007 majority HDDs in static storage with some mobile applications
Media Contents @ Home

- Full-HD movie (MPEG2, 1hr) → 10GB

(source: Storing Your Life – Customer Digital Storage, T. Coughlin, 2008)
1. Data Growth @ Home

2. Analysis of Our Digital Home

3. Analysis of Existing Home Storage

4. Summary & Future Work
Analyzing Our Digital Home!

In-home digital devices

In-home storage devices

Interactions
Networked Digital Home Devices

- Classified into **fixed** -or- **mobile**
- Most, connected to **Home Network** (WiFi)
- All, providing **USB** ports (host-fixed, device-mobile)
- Most, **not allowing s/w changes** within devices

<table>
<thead>
<tr>
<th>Device</th>
<th>Type</th>
<th>WiFi</th>
<th>Ethernet</th>
<th>USB</th>
<th>S/W change</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC/Laptop</td>
<td>Fixed, Mobile</td>
<td>O</td>
<td>O</td>
<td>Host</td>
<td>O</td>
</tr>
<tr>
<td>DTV</td>
<td>Fixed</td>
<td>-</td>
<td>-</td>
<td>Host</td>
<td>-</td>
</tr>
<tr>
<td>DVR(TiVo)</td>
<td>Fixed</td>
<td>O</td>
<td>O</td>
<td>Host</td>
<td>-</td>
</tr>
<tr>
<td>Digital camera</td>
<td>Mobile</td>
<td>O</td>
<td>-</td>
<td>Device</td>
<td>-</td>
</tr>
<tr>
<td>Mobile devices (iPad/iPod/android)</td>
<td>Mobile</td>
<td>O</td>
<td>-</td>
<td>Device</td>
<td>O (app only)</td>
</tr>
<tr>
<td>Game players(Xbox)</td>
<td>Fixed</td>
<td>O</td>
<td>O</td>
<td>Host</td>
<td>-</td>
</tr>
</tbody>
</table>
Equipped w/ *Embedded* Storage

- In-home digital devices are mostly equipped with Embedded Storage (**HDD, NAND-Flash**)
Data Attributes @ Home

• Data are created(synced) in different devices
• Then, duplicated(copied) over many devices
Data Attributes @ Home

• Some requires **high availability** with **long-term preservation (+50yrs)**
  
  – family-photos(jpg)
  – family-videos(mpg)
  
  – sizes become **very large**
    (as data quality increases; 1hr-fullHD→10GB+, ultraHD→0.1TB)
  
  – challenge: how to classify data by importance (keep all photos/videos?)

*(source: Why We Need Whole Home Storage Architecture, Intel)*
Home NAS – *centralized storage*

- **Can it be our solution?**
  - budget problem (will pay $324.99 & extra energy bill?)
  - need to understand complicated spec/setup (many technical buzzzzzz words – even for me)
Home storage should be built on top of distributed in-home digital devices

- Should not rely on centralized storage appliance (due to its purchasing cost & management overheads)
Analyzing Our Digital Home!

In-home digital devices

In-home storage devices
Inter-device Data Operations(1)

Mobile-to-Fixed devices

- Connection: **USB-cable** (sometimes WiFi)
- Operation: **data sync** (via dedicated applications)

**Data sync ("the entire dataset/subset")**
Inter-device Data Operations (1)

*Mobile-to-Fixed* devices

- **Data sync** distributes data over multiple devices
- It will make data indexing harder

Data in mobile devices are stored temporarily
• **Sync data should reside in well-known places** (reducing our indexing/searching efforts)

• **Should use limited # of well-known storage locations** (not centralized storage; but minimizing sync data distribution)

• **Well-known storage locations mean embedded storage provided by in-home digital devices**

• **Should allow each data-sync to use dedicated applications**

[r2]. Efficient data-sync platform “for non-painful subsequent data indexing”
Inter-device Data Operations (2)

*Fixed-to-Fixed* devices

- Connection: **USB-flash** or **USB-hdd (WiFi)**
- Operation: *data sharing* in ad-hoc manners *(file copy)*

Data sharing ("data or a group of data")
Inter-device Data Operations(2)

*Fixed-to-Fixed* devices

- Data sharing ≈ indexing + replication (3x)
- Difficult to find sync’ed data
- Ad-hoc data sharing makes many duplicates

```
<table>
<thead>
<tr>
<th>Fixed devices</th>
<th>USB-flash</th>
<th>USB-hdd</th>
</tr>
</thead>
<tbody>
<tr>
<td>data copy</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>fixed devices</td>
<td>sync’ed data</td>
</tr>
<tr>
<td></td>
<td>data copy</td>
<td></td>
</tr>
</tbody>
</table>
```

“Selective” Data sharing
Home storage should be built on top of distributed storage devices. It should provide an efficient data sync platform for easy data findings.

Seamless data sharing (as if data were stored in local disks)

- Users should minimally (not) be aware that they are accessing the remote data (of somebody else).
- Accessing shared data should be as fast as data access to local disks.
- Should minimally (not) have duplicated data (higher data availability may demand data replications).
Regularly Doing Data Backup?

“Do something -before- my disk gets dead”

– most households like to think ‘their digital data are safe forever’ before encountering disaster (lost of the entire family history-photos)

Not at all -or- Occasionally doing it manually...

solution 1: windows backup S/W → don’t know how to use!
solution 2: burn CD/DVD → too many CD/DVD (2TB=400DVD)
solution 3: online backup → too slow to endure presently!
solution 4: copy whole disk to ext USB → manual work!
Storing Data for Grand\textsuperscript{2}child?

“High availability -and- long-term preservation”

No -or- Even haven’t thought about it!

“Uhm... how about printing out all jpg photos?”
(actually, my parents are doing this...$$$$$)

Challenge: how to classify data by importance?
(All photos/videos are NOT equally important...)
Checking Available Disk Space?

“Do something -before- my disk becomes full”

– most households like to think ‘their storage space is unlimited’ before encountering ‘out of space’

Sometimes, but no plan-ahead actions!

solution 1: erase files(data/programs) → losing valuable ones
solution 2: add a new disk → don’t know how!
**solution 3: move files into ext USB → manual work!**
Fully Automatic Management?

“Automation can simplify storage management; peoples in household are bothered by it”

“...Home users (like sys-admin) insist on understanding & being able to affect the decisions made...”

• Should provide assistive (not full-automatic) storage management service
  (peoples are bothered when they don’t know what’s going on)

• Efficient architecture (outsourced storage) for attribute (high availability) & backup management ➔ how to classify data by importance (backup all photos/videos?)

• Should optimize storage usage(via dedupe) & predict upcoming capacity shortage ➔ removing duplicate copies of photos + on-line expansion
“Do anything -for others- Not to find private data”

– In household, all trusted members (sometimes not)
– User accounts → not proper (due to its inconvenience)
– love to share a single account → sometimes ask some privacy(hiding data); not data encryption

Yes, but in an ad-hoc manner...

solution 1: user account for each member → inconvenient!
solution 2: hide them in a deeper directory → sometimes lost!
solution 3: store them in private storage → extra device!

Home storage should be built on top of distributed storage devices. It should provide an efficient data sync platform for easy data finding. It should provide smart data sharing to efficiently find data & reduce duplication. It should provide essential storage management (attribute/backup/capacity).

Privacy for each family member when needed, not always.

- Normally, allows any access to data from all members.
- On request, should provide an easy way to securely hide his/her data from others (household members).
  (data hiding differs from data encryption; don’t expose even the name of data)
Home Storage Requirements

[r1]. Exploiting distributed in-home digital devices

[r2]. Efficient data sync platform (confining sync-locations)

[r3]. Seamless data sharing (like local disks)

[r4]. Assistive data management (for non-technical admin)

[r5]. Selective data access control (data hiding sometimes)

[r6]. Don’t ask sw/hw changes for existing home devices

[r7]. Solution should be intuitive & simple (like TVs)
Talk Outline

1. Data Growth @ Home
2. Analysis of Our Digital Home
3. Analysis of Existing Home Storage
4. Summary & Future Work
Why Not Well-Known Solutions?

• **Home NAS?**
  – management overheads (setup)
  – costs & energy (always-on)

• **Distributed file systems (Hadoop FS)?**
  – difficult to install/maintain it (difficult even for technical persons)

• **Cloud Storage (Amazon S3)?**
  – presently, costs & network speeds do matter
Home Storage Solutions

Each of them is *partially satisfying* the home storage requirements

- UofW’s HomeViews [Geambasu’07]
- Whole Home Storage [Intel’09]
- CMU’s Perspective [Salmon’09]
- Microsoft’s Family Accounts [Egelman’08]
- Virtual USB Drive [Nam’08,Nam’10]
HomeViews [Geambasu’07]

• **P2P m/w** for personal data sharing applications
• HomeViews helps applications (on PCs/laptops)
  – create **views** to organize files into **dynamic collections**
  – share **views** in a protected (capability-based) way with others
  – seamless access to remote views (data) like local data

Whole Home Storage [Intel’09]

• A single unified namespace (data sharing)
  – view across all data is accessible (identical) from any connected device

• Storage-level solution
  – keep using favorite app’s (freely move to new app’s)
  – work with existing already deployed PCs & emerging standards (DLNA, CIFS/SMB)

• Access control (privacy)
  – read-only access or read/write access for sharing

(source: A Consumer’s Eye View of Whole Home Storage, Intel)
Whole Home Storage [Intel’09]

• **A unified directory**
  – distributed data on multiple devices
  – accessed from any devices networked in home

• **Benefits**
  – easy picture finding
  – shared download directory
  – easy new system integration

(source: A Consumer’s Eye View of Whole Home Storage, Intel)
CMU’s Perspective [Salmon’09]

- Peer-to-peer architecture
- **View** concept
  - concise description of the data stored on a given device
  - each view describes a particular set of data

- **View-based management** (easy to manage)
  - semantic naming for management
  - more # of users could complete given management tasks correctly than traditional hierarchical systems

CMU’s Perspective [Salmon’09]

- View manager GUI
  - placing replicas
  - crash(backup) mgnt
  - space(capacity) mgnt

- Working on
  - laptops, PCs
  - TiVo(DVR+)
  - file systems of Linux, OS X

(source: Salmon et al., “Perspective: semar...)
MS’s Home Accounts [Egelman’08]

• **Current file sharing model** – hierarchical
  – user’s personal directories are at the top of the hierarchy
  – sharing directories are underneath
  – files/settings are private by default

• **Family accounts system** (privacy)
  – shared files/resources are at the top of the hierarchy
  – personal folders are at the bottom
  – files/settings are shared by default; can be private if a user takes additional action

• **Profile manager** : family profile + personal profiles

MS’s Home Accounts [Egelman’08]

- Prototype’s implemented under Windows XP
- Profile Manager application
  - used to switch between profiles (family ⇔ personal)

Virtual USB Drive [Nam’08]

Exactly same as USB flash memory, but replacing NAND flash -with- distributed network storage

⇒ So, each device employs large-size shared USB disk

"Easily work with any CE devices"

(source: Nam et al., “Prototyping a virtual USB drive...,” ICCCS’08, Daegu University, Nov. 2008)
Virtual USB Drive [Nam’08]

• Its prototype
  – ARM9-based MCU, USB1.0 target, WLAN(11/54Mbps)
  – embedded Linux

(source : Nam et al., “Prototyping a virtual USB drive..,” ICCCS’08, Daegu University, Nov. 2008)
Virtual USB Drive [Nam’08]

• Its architecture
  – USB device driver, iSCSI-enabled network stack
  – seamless USB/iSCSI module (user/kernel-level)

(source: Nam et al., “Prototyping a virtual USB drive...,” ICCCS’08, Daegu University, Nov. 2008)
Cost-aware Virtual USB Drive
(Extended version of Virtual USB, under prototyping)

- Store data(block) onto cloud storage (Amazon S3)
- Cost-aware block mapping

Talk Outline

1. Data Growth @ Home
2. Analysis of Our Digital Home
3. Analysis of Existing Home Storage
4. Summary & Future Work
Summary of My Talk

- Introduced Home Storage Requirements(7)
  - [r1]. Exploiting distributed in-home digital devices
  - [r2]. Efficient data sync platform (confining sync-locations)
  - [r3]. Seamless data sharing (like local disks)
  - [r4]. Assistive data management (for non-technical admin)
  - [r5]. Selective data access control (data hiding sometimes)
  - [r6]. Don’t ask sw/hw changes for existing home devices
  - [r7]. Solution should be intuitive & simple (like TVs)
Summary of My Talk

- Yet, effective **storage** solutions are *not* available
- **Virtual USB drive** can be a good candidate

<table>
<thead>
<tr>
<th></th>
<th>UofW’s HomeViews</th>
<th>Intel’s Whole HS</th>
<th>CMU’s Perspective</th>
<th>Virtual USB</th>
</tr>
</thead>
<tbody>
<tr>
<td>[r1]. distributed</td>
<td>work w/ PCs, laptops only</td>
<td>work w/ fixed devices only</td>
<td>work w/ Linux &amp; Mac OS X</td>
<td>any device (w/ USB host)</td>
</tr>
<tr>
<td>[r2]. data sync</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>synced through virtual USB</td>
</tr>
<tr>
<td>[r3]. data sharing</td>
<td>shared views(files)</td>
<td>single unified directory tree</td>
<td>shared views(files)</td>
<td>USB-interfacing big shared-disk</td>
</tr>
<tr>
<td>[r4]. assistive mgnt</td>
<td>-</td>
<td>-</td>
<td>backup(copies) /space mgnt</td>
<td>-</td>
</tr>
<tr>
<td>[r5]. access control</td>
<td>capability-based</td>
<td>read-only + read/write</td>
<td>-</td>
<td>shared only in big shared-disk</td>
</tr>
<tr>
<td>[r6]. CE sw change</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>no sw change (CE) devices</td>
</tr>
<tr>
<td>[r7]. intuitiv/simple</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>simple</td>
</tr>
</tbody>
</table>
Our Future Work

• **Virtual USB drive**, a building block for smart HS
  – simple concept; working with all in-home devices

• Enhancing “Virtual USB drive” features
  – data consistency with multiple virtual USB drives
  – adding new features: availability + deduplication
  – USB dongle → app on mobile devices (iPad/Android)

• Prototyping cost-aware “Virtual USB drive” (2011.5)
  – initially connecting with Amazon S3
  – integrating with other application (backup)
Future Directions for Smart HS

• For home storage requirements
  – our requirements can be a good start for discussion
  – industry-academy joint works are necessary

• For home storage design
  – based on simplicity (don’t ask too much from family & CE manufacturers)
  – better have home storage reference model & use-cases
  – new feature/performance metrics for evaluations

• Working Group for home storage architecture
Questions & Answering!

Towards Smart Digital Home Storage

Young Jin Nam (youngjin@cs.umn.edu)
Office: Keller Hall 6-196