The Continuing Metamorphosis Of the Web

Dr. Alfred Z. Spector
VP, Research and Special Initiatives
Google, Inc.
WWW 2009, Madrid, 24 April 2009

Abstract

The Continuing Metamorphosis of the Web

The invention of HTML and HTTP catalyzed a path of enormous innovation that was hard to foresee in the early 1990's. The Web's continuing metamorphosis has led to fantastically increased capabilities and economic value. It has catalyzed the creation of distributed systems orders of magnitude larger than any previously built, new programming and distribution models for computer applications, great advances in the fields of information retrieval, entirely new domains for theoretical computer science, and more.

This greatly enhanced web is changing the entire environment and enabling some early research promises to become a reality for most Internet users. In this presentation, I will discuss such examples, and in particular, what happens when speech, image processing, human language translation, and mobility are woven into all we do. I will also extrapolate from some current research and advanced web technologies to paint a picture of the web five-to-ten years out. This should have implications for the computer science community, as well as the vast community that is leveraging the web for ever greater goals.

Outline

1. Federation, Reach, and Evolution
2. Extraordinary Achievements of Note
3. The Evolutionary Path Forward
4. 3 Major Extraordinary Advances Brewing:
   A. Totally Transparent Processing
   B. The Rule of Distributed Computing
   C. Hybrid, not Artificial, Intelligence
5. Some Research Challenges
6. Conclusions

Google

Google's Mission and Google Research

Organizing the world's information and
Making it universally accessible and useful.

A research organization optimized for in situ work

Google

Federation, Reach, and Evolution

- The simplicity of the early web standards were genius
  - Federated name space
  - Access (HTTP)
  - Simple data format (HTML)
  - Extensibility!
- Not over-architected in any dimension
- Brilliant omissions (or at least, mostly so ☺)
  - Security
  - Read-write data
  - Semantics
- *Interesting contrast to wide-area file systems work like AFS*

Google

A Semi-Random Walk to Extraordinary Achievements

- The virtuous circle
  - Initial simplicity begat data and usage
  - Usage generated more data and transactions
  - Data modalities diversified
  - User experience blossomed
- Architectural limitations were addressed as needed
- *A bottom-up architectural evolution repeatedly favoring local optimization has resulted in truly momentous results.*
  - The virtual Library of Alexandria
  - The search engine
  - The serving of the long tail
  - Vast changes in business models

Google

Additional, Architectural Achievements

- Network
  - The Web became the catalyst for the rapid internet build-out
- The High Performance Cluster (old word, “multi-computer”)
  - The federated architecture, perhaps strangely, did not obviate the need for large processing complexes.
    - Some workloads require high throughput, low latency, massive data, albeit, embarrassingly parallel
    - The answer: parallel clusters of $O(10^5)$ CPUs & $O(10^{17})$ storage
    - Note “An order of magnitude is a qualitative difference!”
- Browser
  - With a few plug-ins, the application programming model of choice
  - Perhaps, the key client operating system functionality

Google

The Evolutionary Path Forward to New Accomplishments

- Application mix will continue to grow in unpredictable ways:
  - Four areas in flux today: *publishing, education, healthcare and government*
- Systems will evolve: ubiquitous high performance networking, distributed computing, new end-user devices, ...
- Three truly big results brewing:

  1. Totally Transparent Processing

  2. Ideal Distributed Computing

  3. Hybrid, Not Artificial, Intelligence

Google

Totally Transparent Processing

Totally Transparent Processing

$\forall d \in D,\ \forall l \in L,\ \forall m \in M,\ \forall c \in C$

D: The set of all end-user access devices	L: The set of all human languages	M: The set of all modalities	C: The set of all corpora
Personal Computers	Current languages	Text	The normal web
Phone	Historical languages	Image	The deep web
Media Players/Readers	Other forms of human notation	Audio	Periodicals
Telematics	Possible language specialization	Video	Books
Set-top Boxes	Formal languages	Graphics	Catalogs
Appliances	…	Other sensor-based data	Blogs
Health devices		…	Geodata
…			Scientific datasets
			Health data
			…

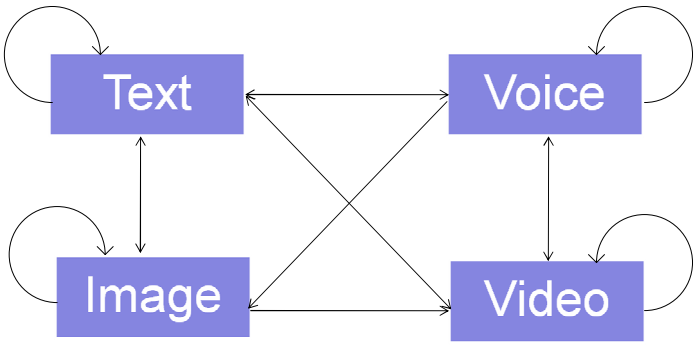
Types of Transparent Processing

- Search, of many forms
- Navigation and Suggestion
- Transformational Communication
- Information Fusion

- Some Google Examples:
  - Universal search
  - Voice Search
  - Find Similar, applied to images
  - Google Translate, particularly in mash-ups
  - Combining images and maps
  - Audio transcription
  - Images and 3d models

Google

Fluidity Among the Modalities



Last two arrows are easily conceivable.

Google

The New Frontier of Web Search – Better/Faster Queries

Query completion before: *Used a fixed dictionary, e.g., in emacs, bash, T9, etc.*

Query suggestion today: *Model queries with query logs, serve them dynamically*

Technical challenges:

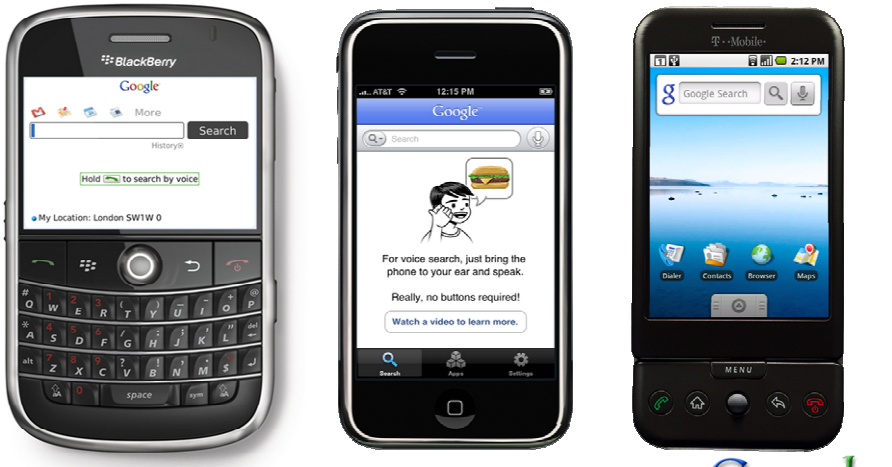
- *response-time, coverage, freshness, corpus dependency (YouTube, image, mobile)*
- *domain dependent:* `rea -> real madrid` *good suggestion in Spain*
- *diversity (danger of popularity), filtering out duplicates, inappropriate results, etc.*

Impact: *Made possible by scale,*

- *the richer the query log corpus, the better*
- *the faster the response time, the better*

Google

Voice Search



Google

Challenges and Rationale for Success

Technically this is very challenging:

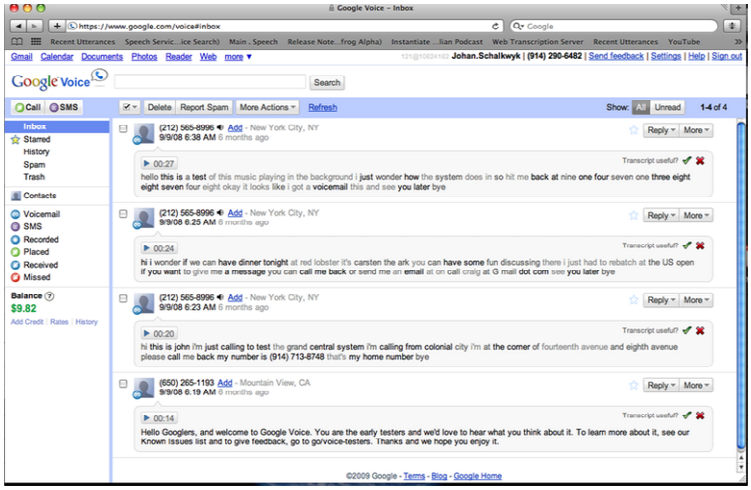
- Huge vocabulary
- Variability in accent
- Background noise

What makes this possible:

- Scalable technology
- Data inputs: Query logs, voice logs
- Compute power

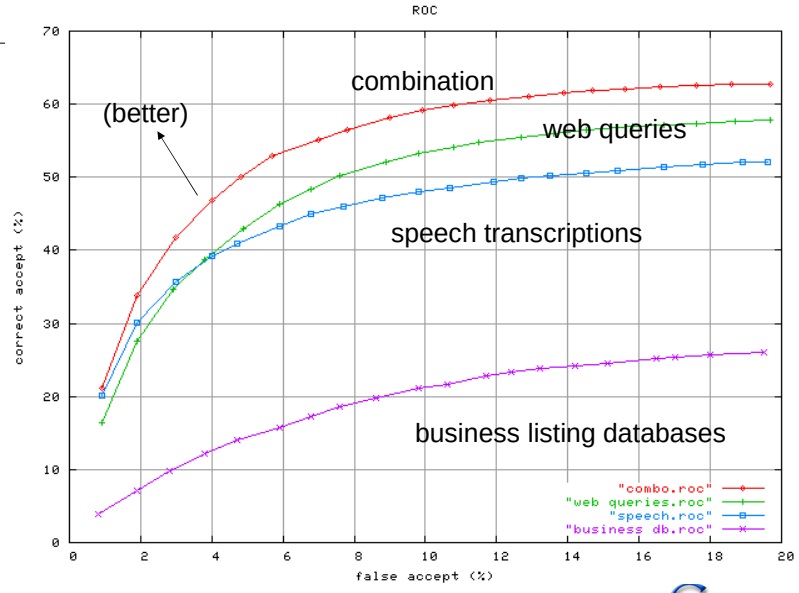
Google

Transcriptions in Google Voice



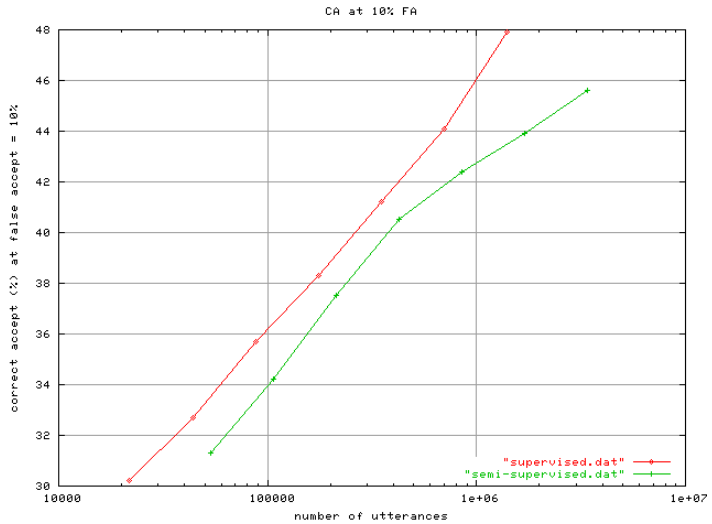
Google

Lots of data: utterance ROC curves: incl. reiection)



Google

The Benefits of Unsupervised Training



Google

Google Translate

***RBMT** – Rules-based machine translation*

***SMT** – Statistical (data-driven) machine translation*

2001-2004
RBMT – 3rd Party
2001
- *French*
- *Italian*
- *Spanish*
- *Portuguese*
- *German*

2004
- *Chinese*
- *Japanese*
- *Korean*

2006
SMT - Google
- *Chinese*
- *Arabic*
- *Russian*

RBMT – 3rd Party
- *French*
- *Italian*
- *Spanish*
- *Portuguese*
- *Japanese*
- *Korean*

2007
SMT - Google
- *Arabic*
- *Chinese (S)*
- *Chinese (T)*
- *Dutch*
- *French*
- *German*
- *Greek*
- *Italian*
- *Korean*
- *Japanese*
- *Russian*
- *Spanish*
- *Portuguese*

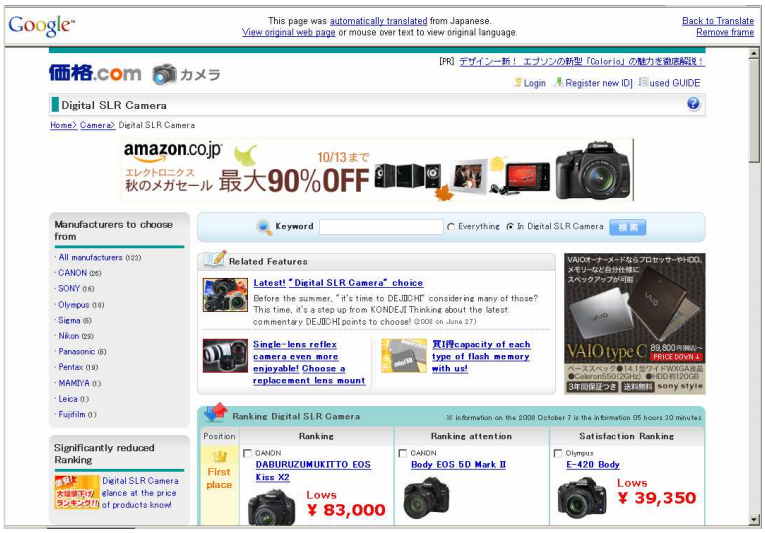
2008
SMT - Google
- *Arabic*
- *Bulgarian*
- *Catalan*
- *Chinese (S)*
- *Chinese (T)*
- *Croatian*
- *Czech*
- *Danish*
- *Dutch*
- *Filipino*
- *Finnish*
- *French*
- *German*
- *Greek*
- *Hebrew*
- *Hindi*
- *Indonesian*
- *Italian*
- *Japanese*
- *Korean*
- *Latvian*
- *Lithuanian*
- *Norwegian*
- *Polish*
- *Portuguese*
- *Romanian*
- *Russian*
- *Serbian*
- *Slovak*
- *Slovenian*
- *Spanish*
- *Swedish*
- *Ukrainian*
- *Vietnamese*

2009
SMT - Google
- *Albanian*
- *Arabic*
- *Bulgarian*
- *Catalan*
- *Chinese (S)*
- *Chinese (T)*
- *Croatian*
- *Czech*
- *Danish*
- *Dutch*
- *Estonian*
- *Filipino*
- *Finnish*
- *French*
- *Galician*
- *German*
- *Greek*
- *Hebrew*
- *Hungarian*
- *Hindi*
- *Indonesian*
- *Italian*
- *Japanese*
- *Korean*
- *Latvian*
- *Lithuanian*
- *Maltese*
- *Norwegian*
- *Polish*
- *Portuguese*
- *Romanian*
- *Russian*
- *Serbian*
- *Slovak*
- *Slovenian*
- *Spanish*
- *Swedish*
- *Thai*
- *Turkish*
- *Ukrainian*
- *Vietnamese*

2001 - 2004 *2006* *2007* *2008* *2009*

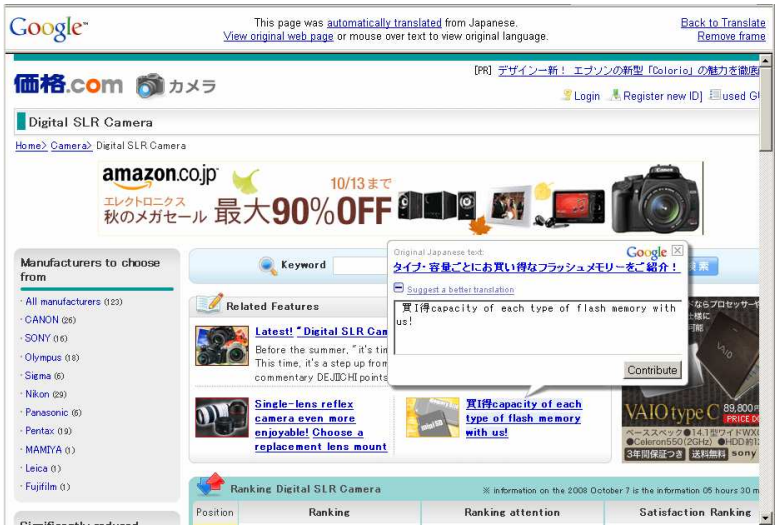
Google

Web Translation



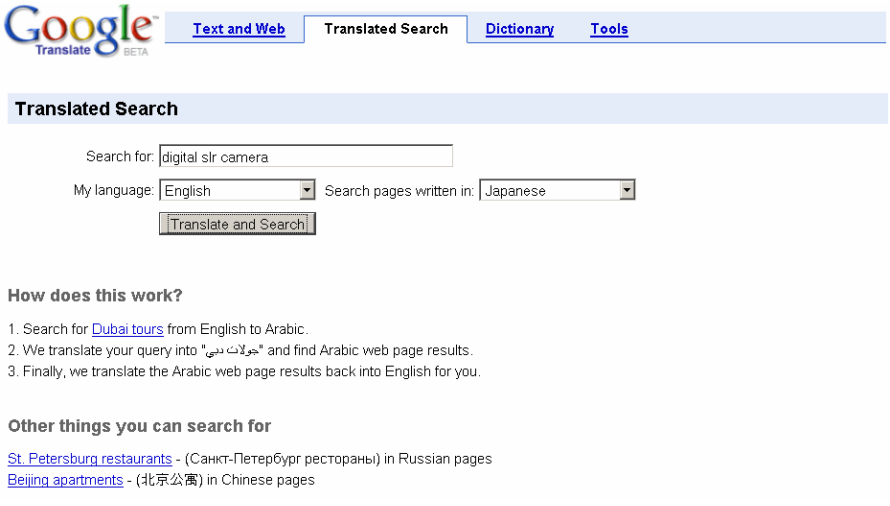
Google

Web Translation Feedback



Google

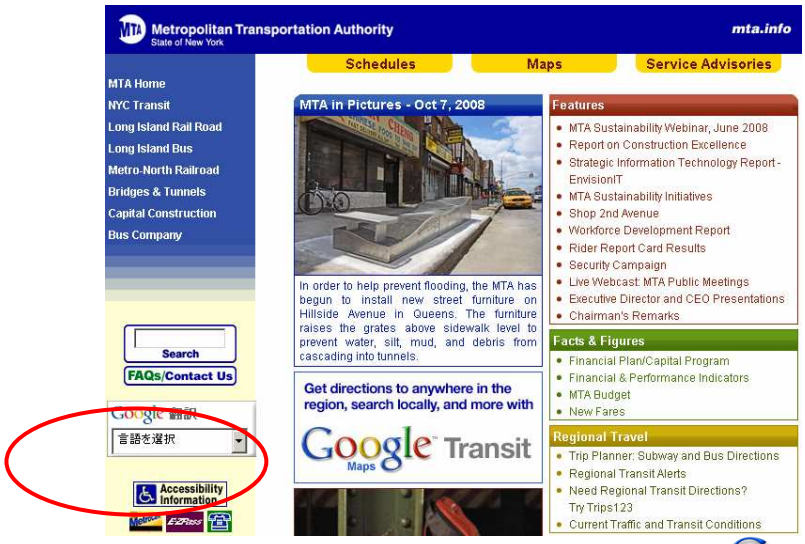
Cross-language search



Google

Cross-language search

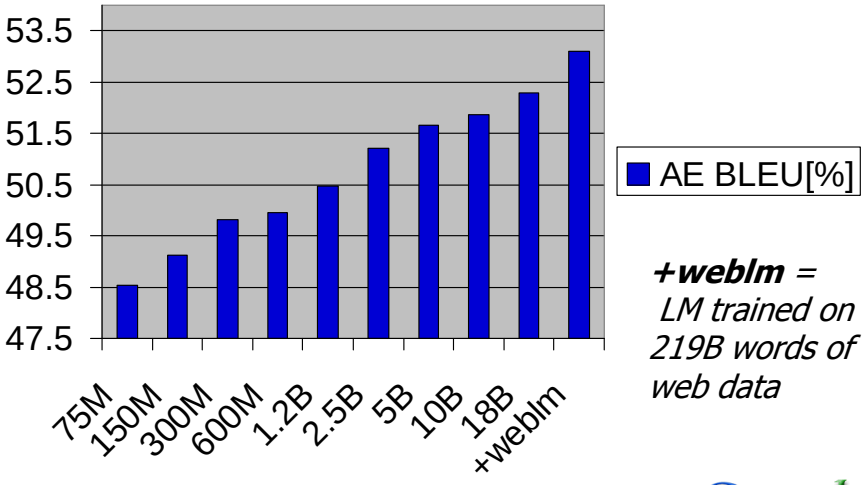
Translate My Page Gadget



YouTube Caption Translation

Google

Impact of data - More data is better …



+webIm = *LM trained on 219B words of web data*

News data

Google

Challenges in Image processing

Visual Semantics

- Recognition (people, landmarks, objects)
- Machine Learning

Correspondence

- Matching images and videos
- Image mosaicing

Geometry

- Ego-motion estimation
- Multi-view stereo

Image Processing

- Maps from aerial imagery
- OCR in all the world's languages

Google

Image Analysis in Image Search

- Image Search helps users find the image they want quickly.
- Understanding the actual content of an image is critical.
- We've been focusing more and more on analyzing images
- This has been rolling out over the last year.
  - Both as user visible filters
  - Behind the scenes in our back-ends.
- Genre filters like clip art / line drawings / color are great examples
    - [flowers], line drawings, clip art, photo, face
    - [porsche] , red, green, yellow, orange, ...

Google

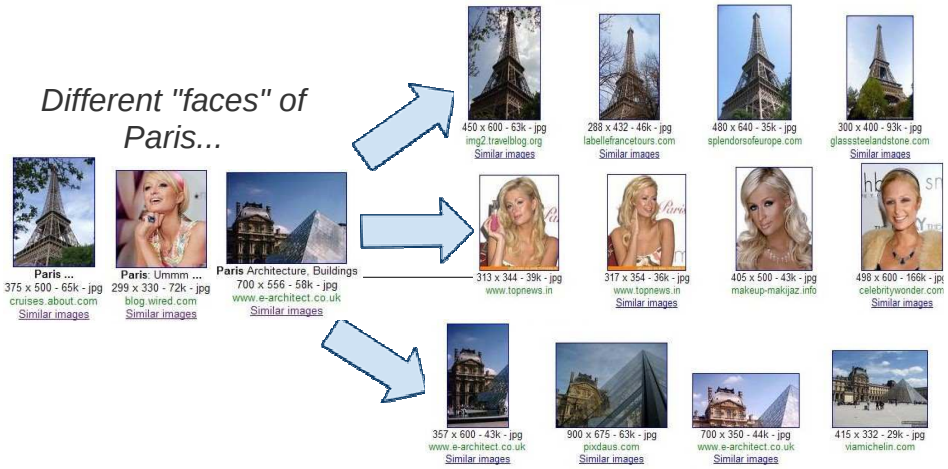
Similar Images in Image Search

- Google has just launched a "Similar Images" feature.
- Accessed by clicking on the similar images link under an image.
- It can also be accessed via preview thumbnails in the result frame.
- We think this will create a major shift in how to search for images.
- Searching for images can now become a navigational experience, where the text (or voice) query acts as a starting point

Google

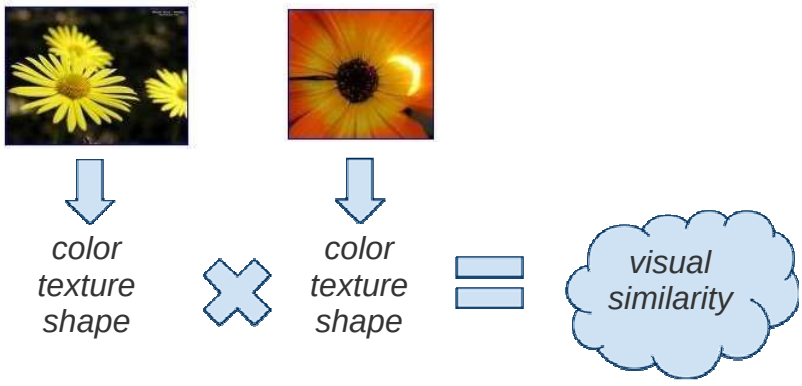
Similar Images in Image Search

Refine by the content of a specific image.



Similar Images in Image Search

- A variety of features are used to determine visual similarity.



Google

Information Fusion across Modalities

Get lists of people names from the web by name detection techniques (NLP)

Scrape image search results using those candidate names as queries

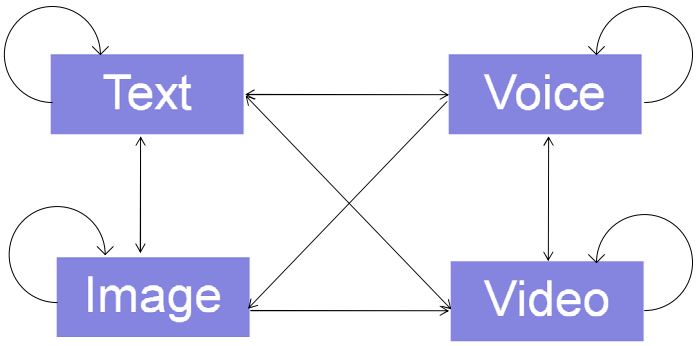
Detect faces in the images and generate face signatures

Apply consistency learning to learn face models

Localize known people in Image/Video

Google

Totally Transparent Processing In-Process…

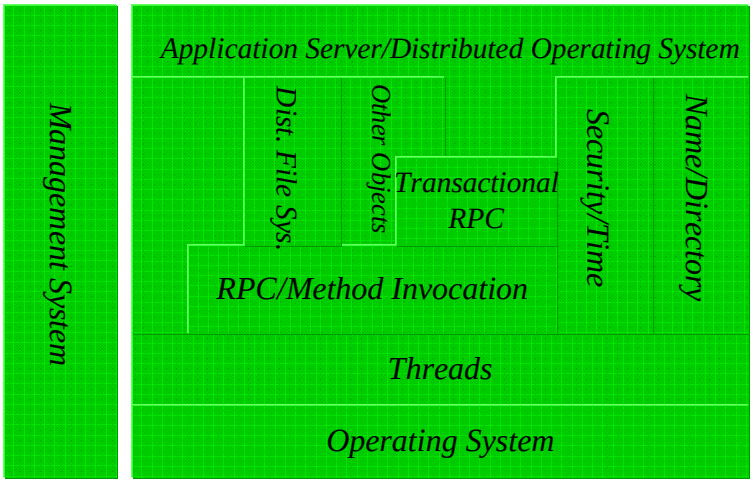


Last two arrows are easily conceivable.

Google

Ideal Distributed Computing

Orthodox Architecture of 70's and 80's



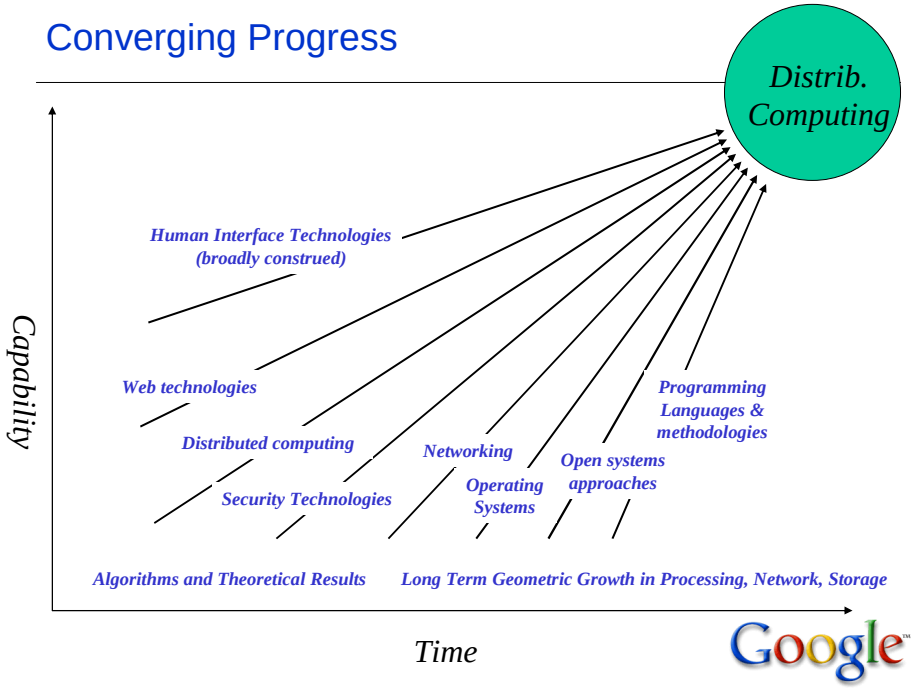
From AZS Pres. To US National Research Council Study on Dependability, May 18, 2004, after a late 80's talk at Univ. of Michigan

What Wasn't Internalized Very Well

- The application mix
- The true nature of global, open systems:
  - Implications on systems, applications, mix and match.
- The implications of *operations* at true scale
  - E.g., work on programming & runtimes predominated system mgmt.
- The complexity of the architecture that would result
  - We tend to assume, *if we can conceive it, it's okay.*
- The collection of further abstractions that would build on fundamentals then known
- In summary there was a limitation of understanding of (truly) large-scale, open integrated distributed systems

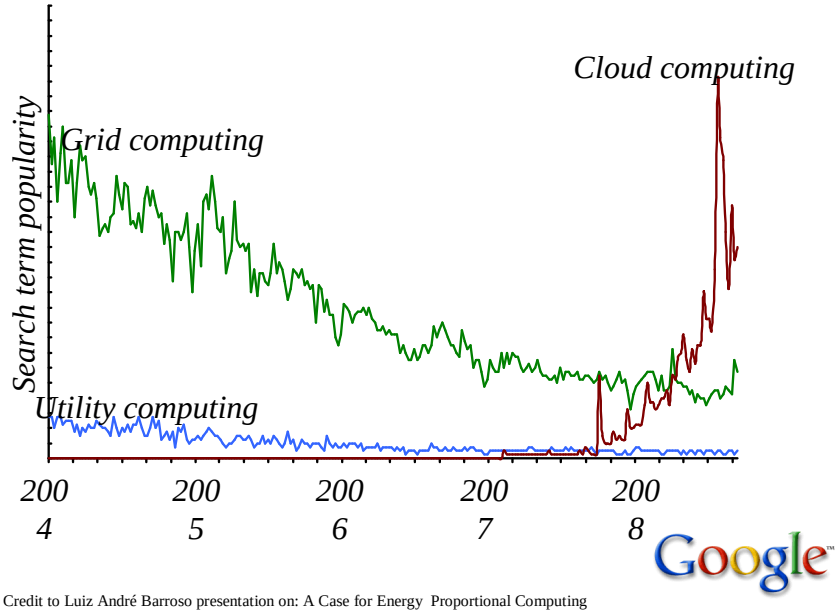
Google

Converging Progress



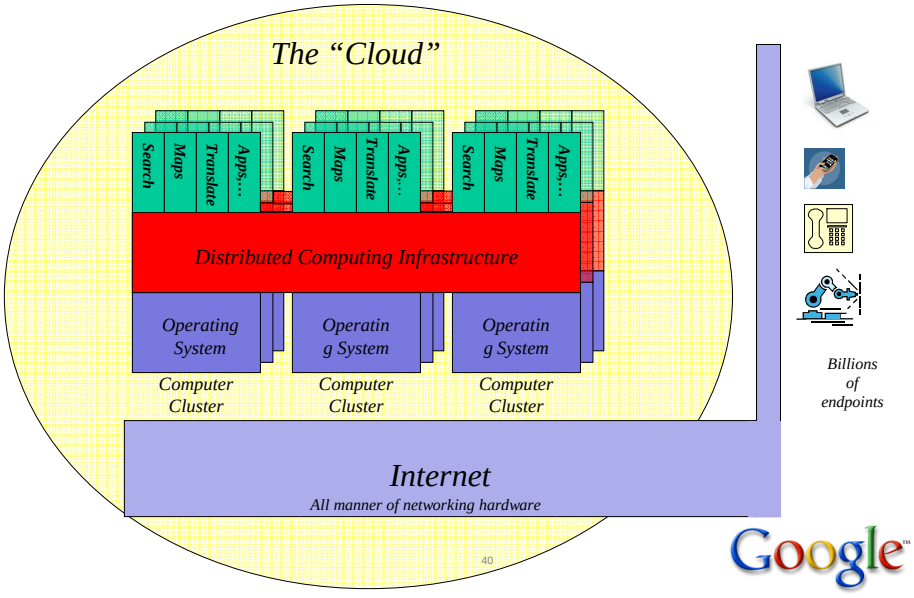
Google

Terminological Evolution



Credit to Luiz André Barroso presentation on: A Case for Energy Proportional Computing

Cloud Computing Architecture



Excitement in Distributed Systems

- Size of user community
- Storage Scale (requiring various characteristics)
  - E.g., security, privacy, availability,
- Processing Scale
  - High performance batch processing
  - High throughput
  - Low latency
- Rapid dynamics
- Highly variable end-user devices
- Communication Scale
  - Bandwidth
  - Endpoints
- Efficiency
  - Equipment
  - Communication
  - Power
  - Management
- Extensibility
- Compliance
- *And more to come, no doubt*

Ideal Distributed Computing

Large networked clusters grow in a fully distributed world

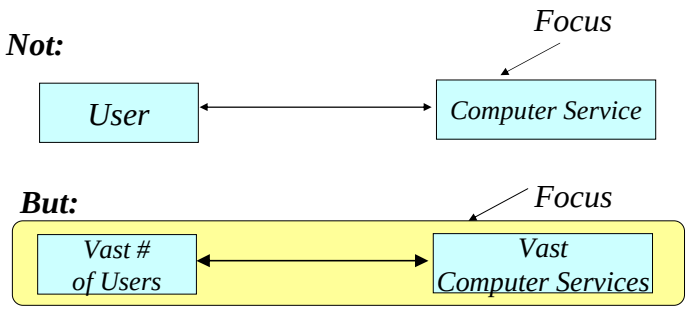
- Arbitrarily high volume transactions
- And, various, partitionable batch process for learning, fusion, etc.
- Network
  - Response-time and bandwidth as needed
- Cluster Processing, or “Cloud Computing” growing ever larger
  - Massive parallelism to hit sweet spot of capital & operating efficiency
- Distributed computing
  - Data sharing, function shipping, as needed
  - Connected and disconnected operation, as seamless as possible
  - Auto balancing of loads between client device and cloud elements
  - Emphasis on manageability (newly, to handle consumers’ many endpoints)
- Significant efficiency gains

Hybrid, Not Artificial, Intelligence

Hybrid, not Artificial, Intelligence

- "Artificial Intelligence" aimed at having computers as capable as people, often in very broad problem domains
- It has proven more useful for computers rather:
  - To extend the capability of people, not in isolation
  - To focus on more specific problem areas
- Aggregation of user responses has proven extremely valuable in learning
- Examples
  - Feedback in Information Retrieval; e.g., in ranking or spelling correction
  - Machine learning; e.g., image content analysis, speech recognition with semi-supervised learning
- Another example of bottom up successes

Key: Holistic Approach To Design



- **Implications**
  - Users and computers doing more than either could individually.
  - Virtuous circle from: ***Data and Processing, Reach, Feedback in a virtuous circle.***

Google

Empiricism - Let Measurement & Feedback Rule

-0.54 *2.7M* *0.55060*

108 milliseconds

$4.78 RPM

-0.0000339 *2,800,000,000 views*

425,440.01 *56.76%* *17.35*

108 seconds/search

9995.55 ***1.3 searches per user***

480,000,000 total pageviews

1607.44 *10,400* *6.55*

$0.303 CPA *108*

$7,660,400

Google

My Long-held View on Semantics, Syntax, & Learning

- Large scale learning has proven surprisingly effective
- Learning is occurring over increasingly variegated features:
  - Both Semantic
  - And Syntactic, and generated in multiple ways
- In my WWW 2002 (Architecting Knowledge Middleware) and Semantic Web 2005 Keynotes, I referred to this as *The Combination Hypothesis*
- Today, I would refine this as the combination of approaches *and* learning from people.

Google

Research Challenges

Google

Challenges in Transparent Computing & Hybrid Intelligence

- Endless applications, with very new user interface implications
- Addressing limits to data
- Techniques to integrate user-feedback in acceptable fashions
- Approaches to new signal (e.g., annotations)
- Explanation, scale, and variance minimization in machine learning
- Information fusion/learning across diverse signals – The Combination Hypothesis, more generally
- Usability: devices and subpopulations

Google

Research Challenges in Ideal Distributed Computing

- Alternative designs that would give better energy efficiency at lower utilization
- Server O.S. design aimed at many highly-connected machines in one building
- Unifying abstractions for exploiting parallelism beyond inter-transaction parallelism and map-reduce
- Latency reduction
- A general model of replication including consistency choices, explained and codified
- Machine learning techniques applied to monitoring/controlling such systems
- Automatic, dynamic world-wide placement of data & computation to minimize latency and/or cost, given constraints on
- Building retrieval systems that efficiently and usably deal with ACLs
- The user interface to the user's diverse processing and state

Google

3 Interesting Challenges

- Security and Privacy Technologies and Policy
- Application of technologies to health
- Applications to Government

Google

Conclusions

- The Web's brilliant initial design lead to a series of local optimizations with extraordinary results
- Evolutionary advances continuing
- They are aggregating into at least 3 major advances:
  A. Totally Transparent Processing
  B. The Rule of Distributed Computing
  C. Hybrid, not Artificial, Intelligence
- Challenges for academics and industrial researchers/engineers abound

Google

¡Muchas Gracias!

Thank you very much!