In the mid-80’s Joel Birnbaum of HP Labs first talked about pervasive/utility computing. In the second half of the 90’s HP Labs began work on “e-services computing.” Anyone remember e-speak? In the late 90’s HP Labs began work on planetary scale computing. Invent a utility that allocates globally connected resources on demand, anytime, anywhere. Realize service centric computing.

Planetary scale computing

- A new computing model that allocates IT resources on demand, anywhere
  - Static or dynamic
  - Economical
    - Commodity computing
    - Self-aware with proactive control
  - Programmatically configured
    - Rather than re-cabling...
    - Federated “bricks” of server and storage
  - Federating on a planetary (geographic) scale
    - $O(10^5)$ element data centers

IT infrastructure becomes a virtual resource service
- Description, specification
- Provisioning
- Management
- Billing
- Trust

Conceptual target:
- thousands of resources per service, thousand of services per data center, thousands of data centers
Resulting challenge

- "pay as you grow" is attractive
- horizontal scaling enables it
- managing the sheer number of devices that results (>10k node data centers) is a problem
  - the largest cost in IT is data center operations and management

```
Source: Giga May 2001
```

basic idea: consumer/supplier

intelligent provisioning: effective use of physical resources
intelligent brokering: match service demand with resource capacity

```
Planetary Computing
```

Basic idea (cont.)

Layer 3 Services
- Services consist of application components on virtual resources
- Service semantics / SLAs

Layer 2 Virtual Resources
- Intelligent provisioning of virtual resources to physical service cores
- Resource planning based on QoS/power/cost/geography/administration

Layer 1 Physical Resources
- Servers, storage and switches
- Software images, licenses
- Data center resources: power, floor space, economics

```
```
**Scalable commodity open source platform**

- ia-64 Linux kernel design and implementation

- http://www.gelato.org

---

**Dynamic thermal management in large scale data centers**

- Power Density -
  - Microprocessor: 200 W/cm² (by 2003, today 60 W/cm²)
  - System - 300 W, thin 1U form factor 10 to 15 kW per EIA Rack footprint
  - Room - 2700 W/m² (~300 W/ft²)
- Use 3D modeling to understand thermal characteristics of data centers
- Exploit this for dynamic resource allocation

---

**Planetary scale storage**

- iShadow

- from islands of isolated data to anywhere, anytime access to data

---

**Hippodrome: automatic storage management**

- Design system to meet workload requirements
- Learn workload performance characteristics
- Configure devices & migrate data
- Analyze workload
- Implement design

- Benefit: "autonomic" storage

---

**Dynamic thermal management in large scale data centers**


---

**Planetary scale storage**


---

**Hippodrome: automatic storage management**


---
Self aware services

- How to manage 50,000 servers, 1 million objects?
  - centralized management, human-centered operation, polling architectures don't scale
- services monitor own health and the health of local dependents to determine the root cause of failures
  - based upon statistical measures and bayesian network reasoning

"Self-Aware Services: Using Bayesian Networks for Detecting Anomalies in Internet-based Services"; Bronstein, Alexandre; Cohen, Ira; Das, Jaydeep; Duro, Manoj; Friedrich, Richard; Kleyner, Gary; Mueller, Martin; Singhal, Sharad in Proceedings of Integrated Network Management VII (IM-2001), 14-18 May2001, Seattle, IEEE/IFIP

SmartFrog: service description and deployment

- Configuration description language
  - precise, desired configuration of applications composed of sets of components running across a distributed system
- Service deployment architecture for massive systems
  - realize application description
  - monitor and manage the resulting applications through their lifecycles


web services everywhere

- Internet
  - XML, WSDL, WSFL supporting application platforms
- mechanisms and services for business interactions
  - Biz Router, UDDI, Rater, SLA Manager
- ESP, ESP, ESP
- Enterprise
- MMO, JEE, MQS
- Internet

business verticals:
  - procurement
  - HR
  - utility computing
reusable, core service blocks:
  - services
  - proxies
  - app APIs

exactly-once transactions

Client ↔ Service ↔ Environment
(Travel Agent)     (Car Broker)  (AVIS/Hertz/Alamo)

Contract:
  - idempotence
  - non-blocking
  - testable
From research to reality

- HP announced the Utility Data Center (UDC) Nov 2001
- Based on HP Labs research on adaptive internet data center:
  - ability to direct resources to any application dynamically
  - self healing, policy driven.
  - Open system: Windows, Linux, HP-UX, Sun Solaris

... to create a dynamically configurable utility fabric that can be programmed per service or customer, based on SLAs and demand...

creating a service with the UDC

1. Architect new service:
   - Svc “A”
   - FW
   - LB
   - WEB
   - WEB
   - WEB
   - APP

2. Build a service template:
   - APP
   - FW
   - 1U Linux
   - 2U NT
   - Svc “A” appliance
   - HP-UX
   - LB
   - WEB
   - WEB
   - WEB
   - APP

3. Ignite the service
   - Install apps
   - Discover and apply free resources
   - Specify connectivity
   - Auto-configure network and storage
   - Auto-load OSes

conclusion

- HP focus on service-centric (utility) computing
- self-management research at all layers (for all the known reasons, but also to deal with new dynamism):
  - storage self-management
  - utility data-center resource allocation
  - self-aware services
  - service lifecycle management
  - exactly-once multi-party web service conversations
  - ...