TACS-2005

June 5, 2005

Smart Chip, System & Data Center

Dynamic Provisioning of Power and Cooling from Chip Core to the Cooling Tower

Chandrakant D. Patel

Distinguished Technologist

Hewlett Packard Laboratories

www.hpl.hp.com/research/dca
chandrakant.patel@hp.com

Business Motivation

Customer Data Center Needs

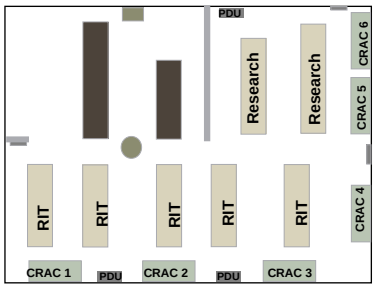
- Compaction & Consolidation
- Improve Uptime
- Lower Operational Costs

State of Art

- Design by intuition and "rules of thumb"
- High energy consumption
- High personnel cost due to complexity and lack of automation
- Inflexible components, over-provisioned

Value Proposition

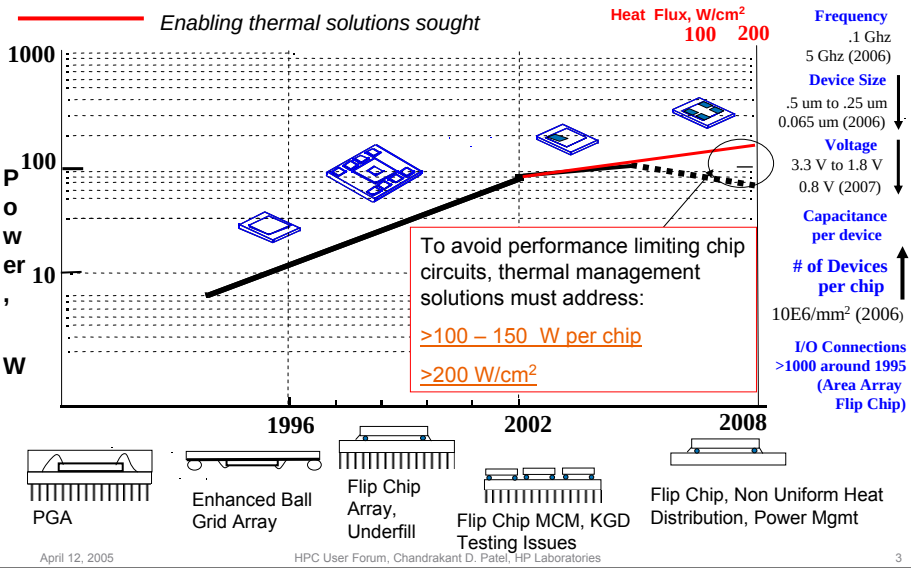
- Innovative capabilities on top of commodity components to create products and services
- Future "smart" data center that adapts to dynamic changes in business processes



High Density Installation

Motivation: Technology Trends

power, flux W/cm², chip packaging, microprocessor organization, semiconductor technology

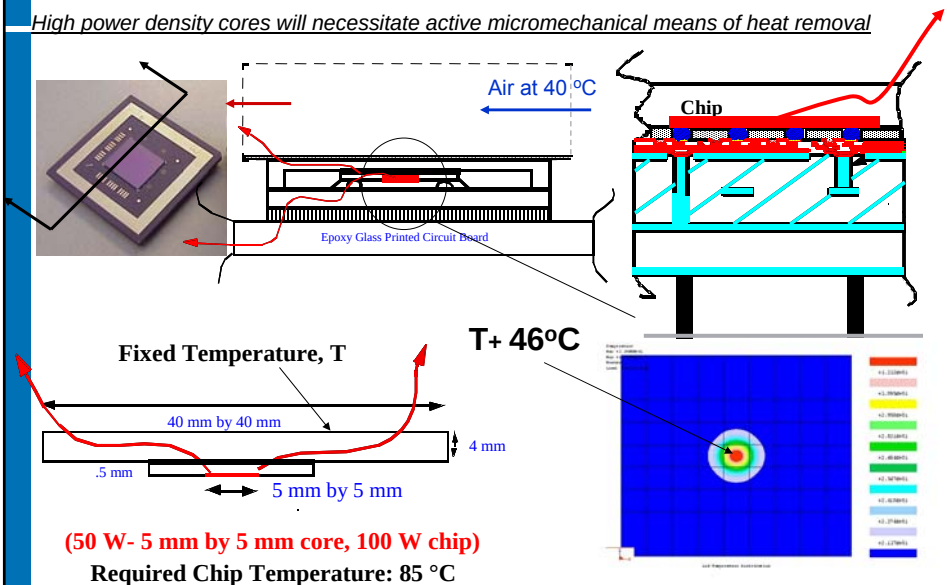


April 12, 2005 HPC User Forum, Chandrakant D. Patel, HP Laboratories 3

Chip

Impact of Small High Power Core(s)

High power density cores will necessitate active micromechanical means of heat removal



(50 W- 5 mm by 5 mm core, 100 W chip)

Required Chip Temperature: 85 °C

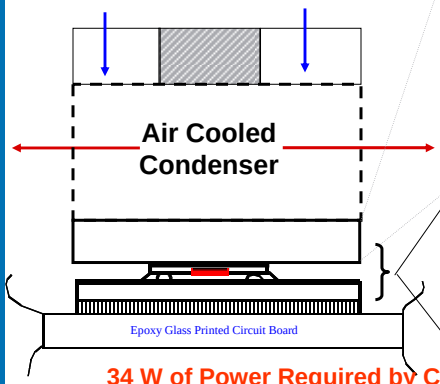
April 12, 2005 HPC User Forum, Chandrakant D. Patel, HP Laboratories 4

Future Integrated Micro-cooler

High Power Density Chips (150 W, 200 W/cm^2)

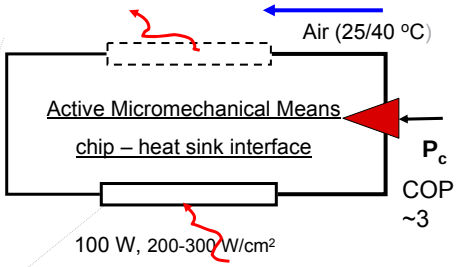
Active Micro-mechanical Cooling at the chip interface, e.g.

- On-chip two phase cooling
- Solid State Active Cooling



Fan Power required: 4 W

~.008 m^3/s, ΔP of 75 Pa; $\zeta_{\text{wire to air}}$: 15%



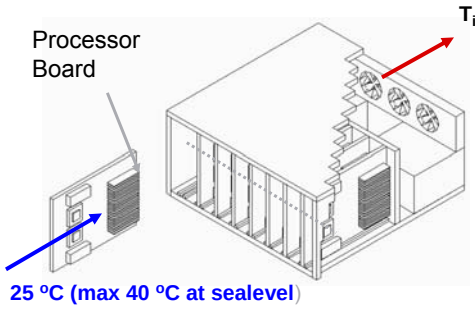
Heat Sink temperature of 70 ºC at an ambient of 40 ºC, which implies:

- *Temperature rise through the interface not to exceed **30 ºC***
- *Active means at interface level*
- **P_c (power required): ~30 W**

34 W of Power Required by Cooling Resources to remove 100 W

Technology Trends - System

Impact of Compaction

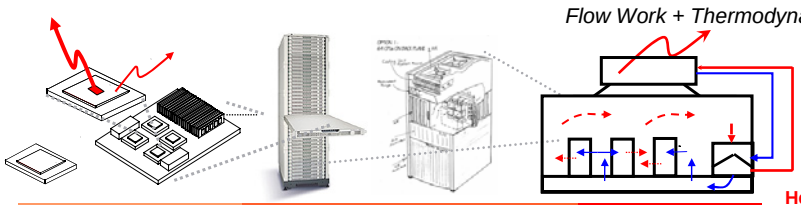


- Single Board Computer: 250 W
- 10X boards per chassis makes a 2.5 KW enclosure;
- Flowrate ~ 150 litres/sec; High pressure drop ~ 200 Pa
- 200W Fan power per enclosure
- **6** enclosures per **rack**; 15 KW rack, 900 litres/sec (~1800 CFM) minimum flow rate **(1.2 KW Fan Power)**

Motivation: Economics & Sustainability

Energy Required to Drive the Global Compute Utility

Recurring Cost Cooling Resources: ***~$1.2 million per annum*** *for a data center with 100 racks of 13 KW each – fully loaded 1U systems in a standard rack*



Flow Work + Thermodynamic Work

5 W **100 W** **15 KW** **20 KW** **1 MW** — **Heat Generated**

10-34 W **1.5KW** **2KW** **1MW** — **Energy to Remove Heat**

Accessing Services

Thin Client Accessing Data Center

10 GW **56 GW** *Global Power Consumption* **6 GW** — **Heat Generated + Energy to Remove Heat**

50 Mega tons **280 Mega tons** **30 Mega tons**

Annualized Destruction of Available Resources – Coal

20 E6 Kg per year NOx

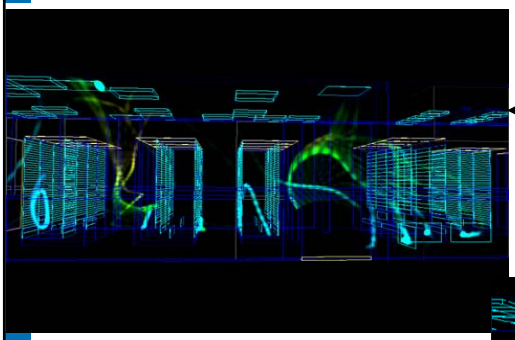
China Summer Shortfall – 30 GW [2]

China : 240 GW [2]

[2]. IEEE Spectrum, 10/2004

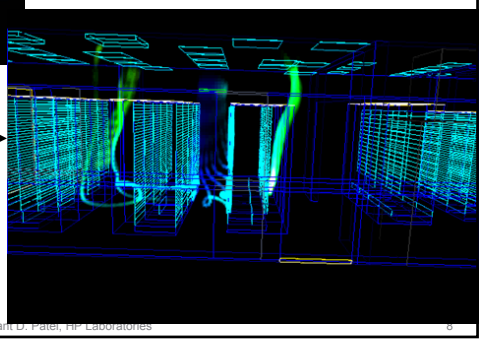
Thermo-fluid Provisioning with CFD modeling

Example of Compaction – **High Density Render Farm**



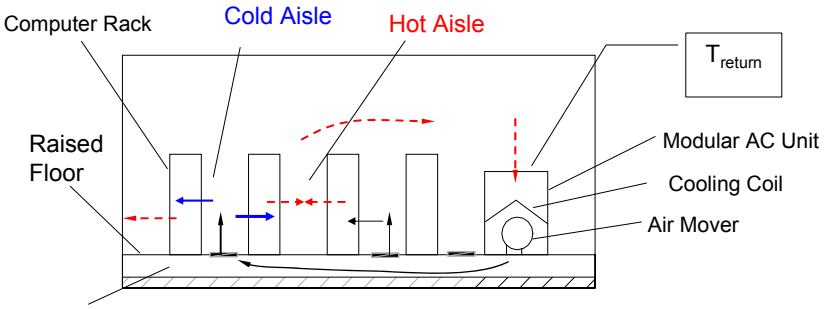
- Excessive residence time of hot air particles
- Energy Inefficient
- Improper provisioning of AC resources

- Reduce resident time of hot air particles
- Minimize temperature
- Proper provisioning of AC resources



Data Center Flow Distribution & Control

State of Art Design and Control



Plenum with cold air return, hatching signifies blockage from cables, piping, etc

Cool Fluid

Hot Fluid

Energy Aware Chip, System & Data Center

Why is this a compelling research problem?

Limitations in the State of Art

Chip

- Designed and provisioned for full 100 W.

System

- System fans are provisioned for maximum system and rack power
  - *Good design practice optimizes fan selection based on pressure drop*
  - *However, fan speed control based on "single" point measurement of a given chip, or minimal local information.*

Data Center

- Cooling resources provisioned based on the full load in the data center
  - *control based on air temperature sensed at a few points, lack of global knowledge, monitoring granularity lacking, policies for control required*

• Opportunity to cut the power consumption by cooling resources

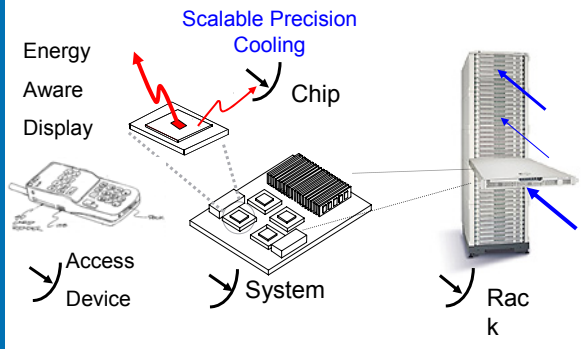
• Target: 50%; ~ 1.5 GW for the data centers alone

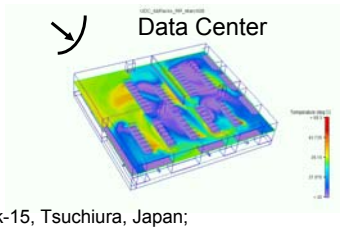
Part 1. Concept: "Smart" Chip, System, Data Center

Dynamic Provisioning of Power and Cooling

Balance of Power & Cooling

• Dynamic allocation of *compute and cooling resources* to minimize energy consumption while meeting user needs





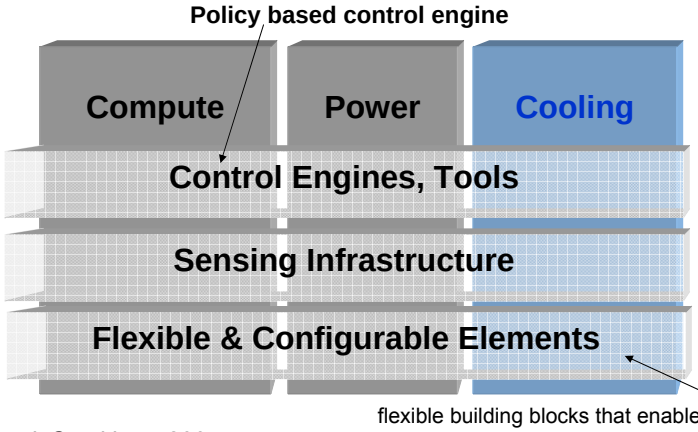
Ref. Patel, Keynote - Energy Aware Computing, 12/2003, ISMME k-15, Tsuchiura, Japan;

Ref. Ranganathan, Energy Aware Display, 2003, Int'l Symposium on Mobile Systems

Smart Chip, System & Data Center

Compute Utility "smart" by design & operation

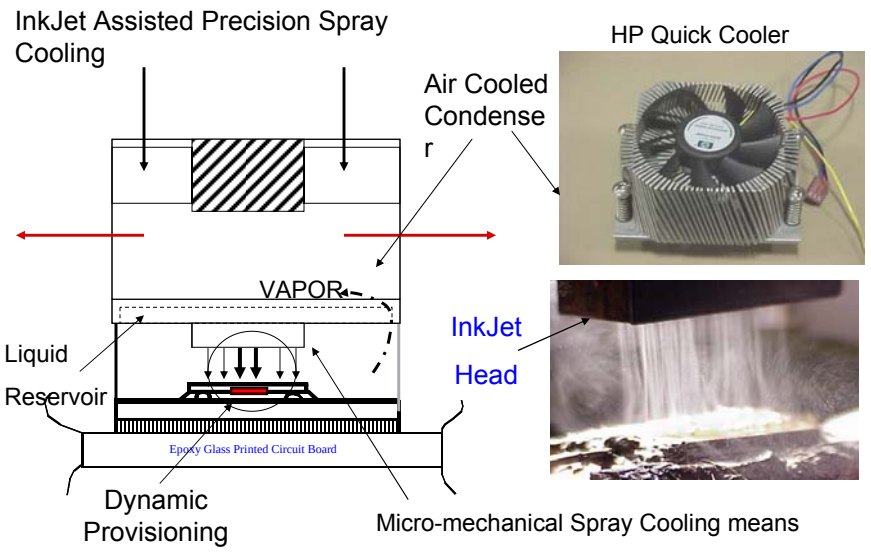
• Physical attributes, sensing, aggregation, logging, policies and metrics for control



* Patel et. al, Semitherm2005

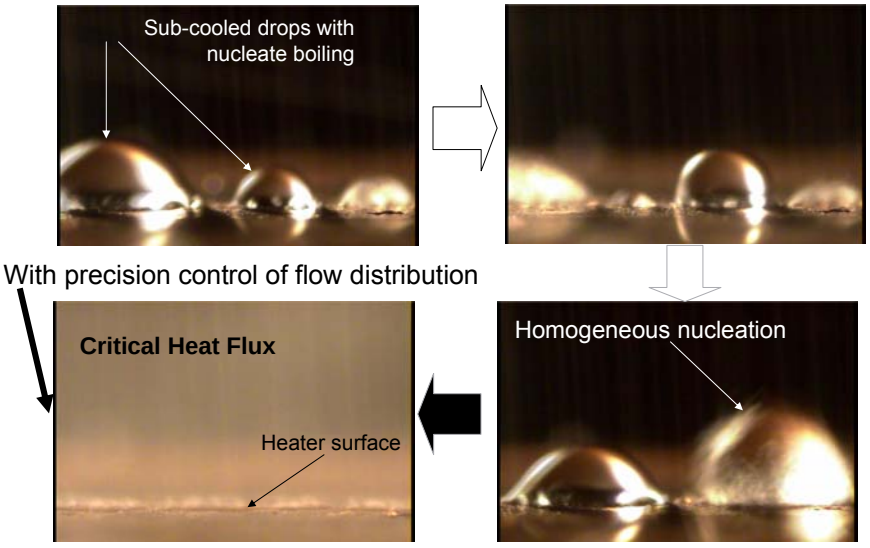
Chip Scale Flexible Building Block

Dynamic Precision Provisioning using Phase Change



Stages of Boiling

Maintaining optimum excess temperature through dynamic spray



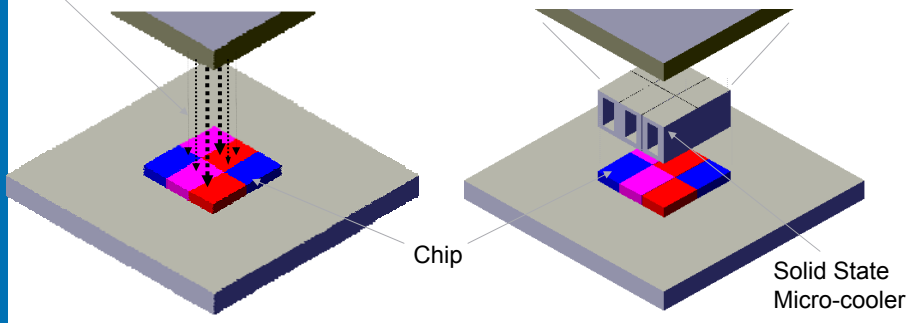
Chip Scale – Power and Cooling

Dynamic Provisioning of Power on the Chip

w/Aaron Wemhoff, U.C. Berkeley, Research Intern at HP Labs, 2003

Provisioned Coolant supply, $m_{region1}$, $m_{region2}$

Condenser/Cold plate



Tuned 2-phase cooler with variable fluid flow rate

Tuned solid state micro-cooler

Colors are varying power/ performance levels of various functionalities based on most efficient available cooling

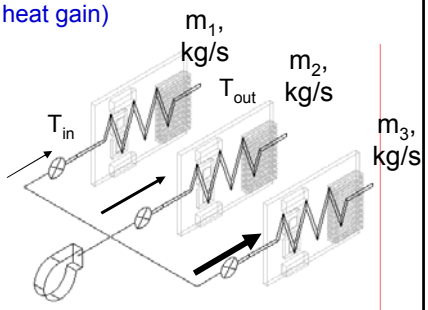
Smart Chip & System

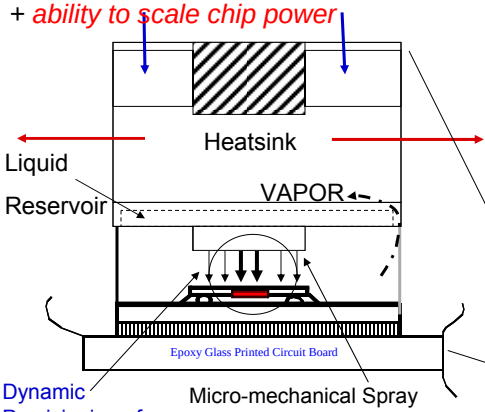
Flexibilities

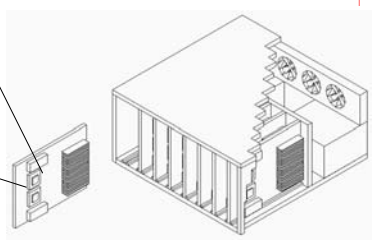
- coolant mass flow (phase change & sensible heat gain)

+ chip level and system level temperature sensing

+ ability to scale chip power







Dynamic Provisioning of coolant

Micro-mechanical Spray Cooling means e.g. InkJet head

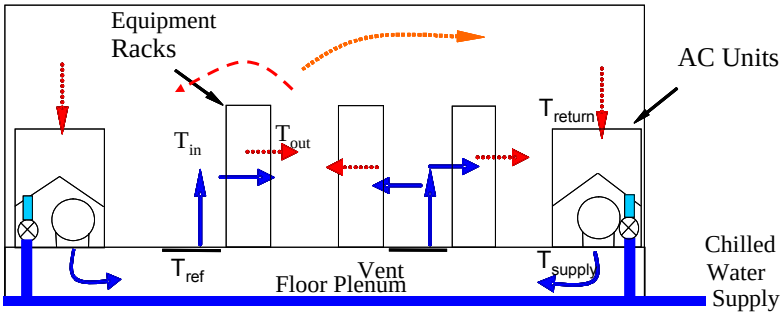
Smart Data Center

Flexibilities - flow work and thermodynamic work

- Coolant mass flow and temperature (thermodynamic)
  - blower speed (flow)
  - adjustable "smart" vent tiles to achieve higher granularity in flow control (flow)
- Coolant temperature – valve opening to change chilled water flow rate

\+ *sensing at system and rack level (100s' of points)*

\+ *ability to allocate power (compute load) at system and rack level*



Thermo-fluids Policy

Quantify the Impact of Air Flow Distribution in the Data Center

Key to maximizing the performance in the data center would be to:

A. Minimize infiltration of hot air into the cold aisles

B. Minimize the mal-provisioning of CRAC units due to mixing

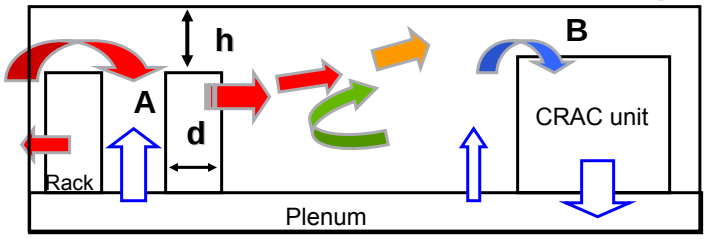
Supply Heat Index (SHI)

- Index of infiltration

Return Heat Index (RHI)

- Index of provisioning

Based on Dimensionless Parameters e.g h/d

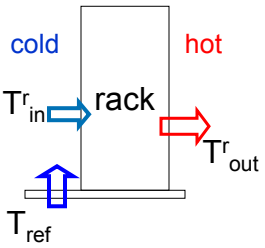


[5] Sharma *et al.*, Dimensionless Parameters, AIAA/ASME 2002

Minimizing the Index of Infiltration

Total Heat dissipation from all racks in data center

$$Q = \sum_j \sum_i m^r_{i,j} C_p \left(\left(T^r_{out}\right)_{i,j} - \left(T^r_{in}\right)_{i,j} \right)$$ [5]



Rise in enthalpy of cold air before entering the racks

$$\delta Q = \sum_j \sum_i m^r_{i,j} C_p \left(\left(T^r_{in}\right)_{i,j} - T_{ref} \right)$$

Supply Heat Index: Ideally 0 [5]

$$SHI = \left(\frac{\delta Q}{Q + \delta Q} \right)$$

Return Heat Index: Ideally 1 [5]

$$RHI = \left(\frac{Q}{Q + \delta Q} \right)$$

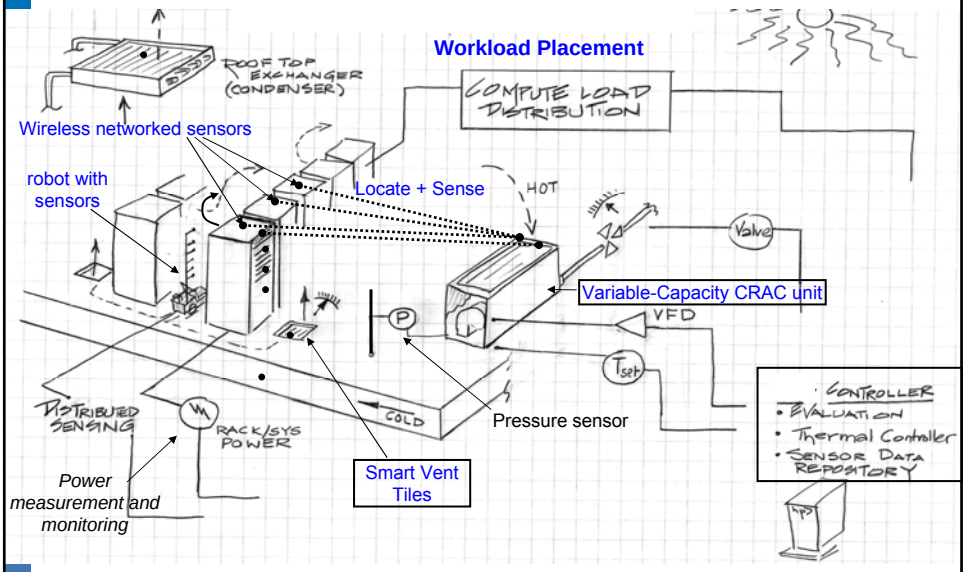
$$SHI + RHI = 1$$

[5] Sharma *et al.*, Dimensionless Parameters, AIAA/ASME 2002

Smart Data Center

Concept

20 ˚C Ext Ambient ⟷ **45 ˚C Ext Ambient**



Implementation of Smart Data Center

HP Laboratories, Palo Alto

- **Ceiling or Room Return**
- **CRAC: Blower and chilled water flow control**
  - Smart Tile – Vent Flow Control

Smart Data Center Realized

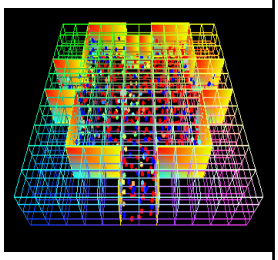
HP Laboratories, Palo Alto, California

+ Wired, Robotic & Wireless Sensors
+ Flexible Infrastructure

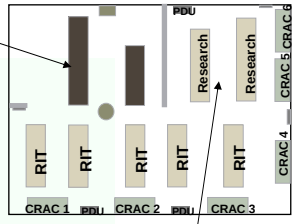
(e.g. Smart Vents)

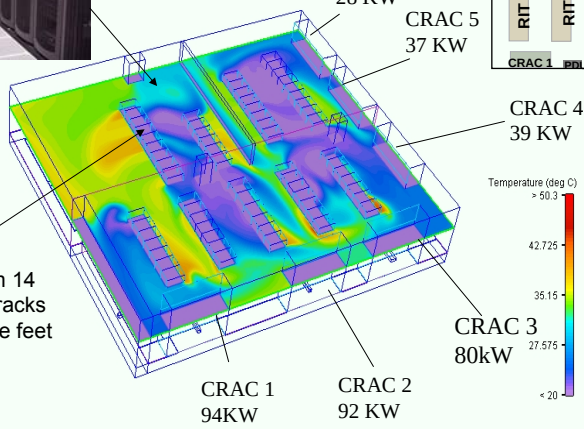
+ Data Aggregation & Logging
+ Visualization
+ Control System

Geo View



Given Production State in the "Smart" Data Center





CRAC 6
28 KW

CRAC 5
37 KW

CRAC 4
39 KW

High Performance Cluster – low load in this state

HD Area:

140 KW from 14 fully loaded racks in 400 square feet

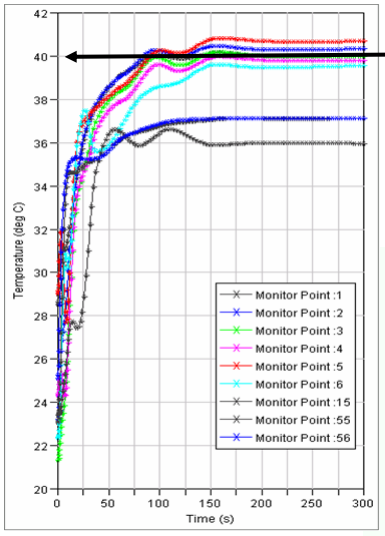
CRAC 3
80kW

CRAC 1
94KW

CRAC 2
92 KW

Static Smart Cooling Analysis

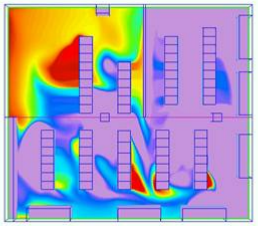
Planning Failure Mitigation – CRAC #1 failure

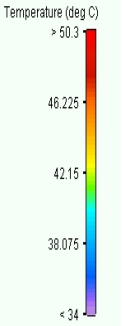


Unacceptable temperature level

Develop Strategies for Failure Mitigation

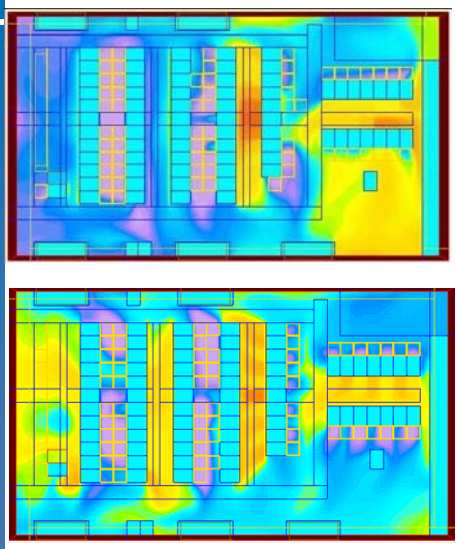
- Workload Migration
- "Smart" Redundancy of AC equipment





** It took 80 seconds to reach unacceptable temperature level.

"Smart" Cooling – Static & Dynamic



Production data center*

* *Data Center Modeling from HP Services*

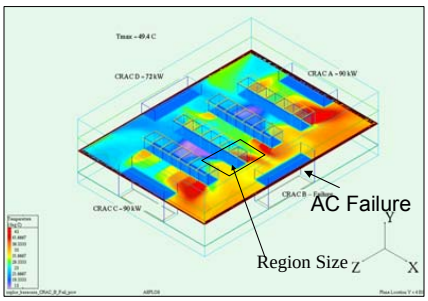
- The Smart Cooling system is a combination of modeling, metrology and intelligent control.
- Two key components:
  - *Static Provisioning of Resources for fixed distribution of resources**
  - *Dynamic Smart Cooling*
- Energy savings with respect to the cooling resources (compared to state of art)
  - up to 25% with static provisioning
  - 50% savings with dynamic – savings of $600K/yr for the data center with 100 - 13 KW racks
  - Target payback ~ 1 year

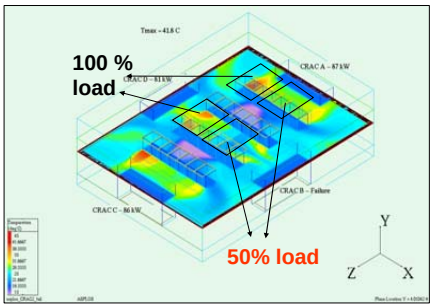
Example of Smart Data Center Operation

power, compute and cooling

- *Temperature-aware resource provisioning* - algorithms to leverage hardware power states to provision heat loads based on thermo-fluids policies
  - *Moore et. al, "Making Scheduling "Cool"…"USENIX 2005*
  - *Sharma et. al, "Balance of Power – Dynamic Thermal Management of IDC, Internet Computing, Jan/Feb 2005*
- Employ workload migration as a knob for resource management

  Janakiraman, et al, "Cruz: Application-transparent distributed checkpoint-restart on standard operating systems," *Dependable Systems and Networks 2005.*
- Dynamic provisioning of cooling resources based on sensing and policy based evaluation of sensed data
  - *Patel et. al, "Smart Cooling of Data Centers" ASME Interpack 2003*
  - *Patel et. al, "Thermal Considerations in Data Center Design", Itherm 2002*





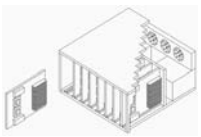
Future: Chip Core to Cooling Tower

Flexible resources, pervasive monitoring, policy based control system

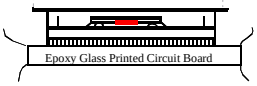
Scalable High level Evaluation Engine – chip to data center stack

Pervasive Monitoring – Distributed Sensing

Information transfer and **policy assignments**



Flexibility in mass flow e.g. coolant with phase change



Flexibility in mass flow of coolant

Flexibility in Air Conditioning resources

- Flow
- Temperature

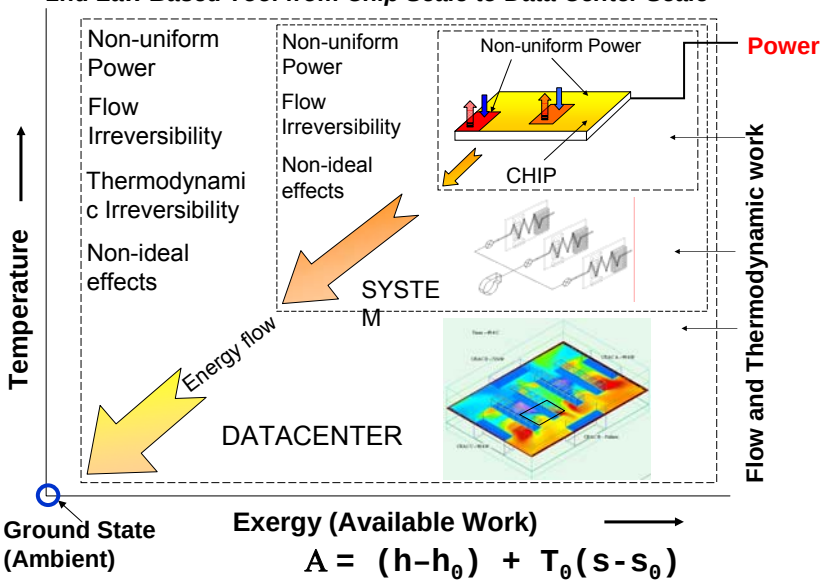
Datacenter

Regional Scalability in Power: Allocate power to a given chip, in a given system, in a given data center in a given locality

High Level Policy based on Exergy

Joint work with U.C.Berkeley – Van Carey and Shah

2nd Law Based Tool from Chip Scale to Data Center Scale



$$A = (h - h_0) + T_0(s - s_0)$$

Metric for Service Delivered

MIPS/Exergy(watts) Destroyed

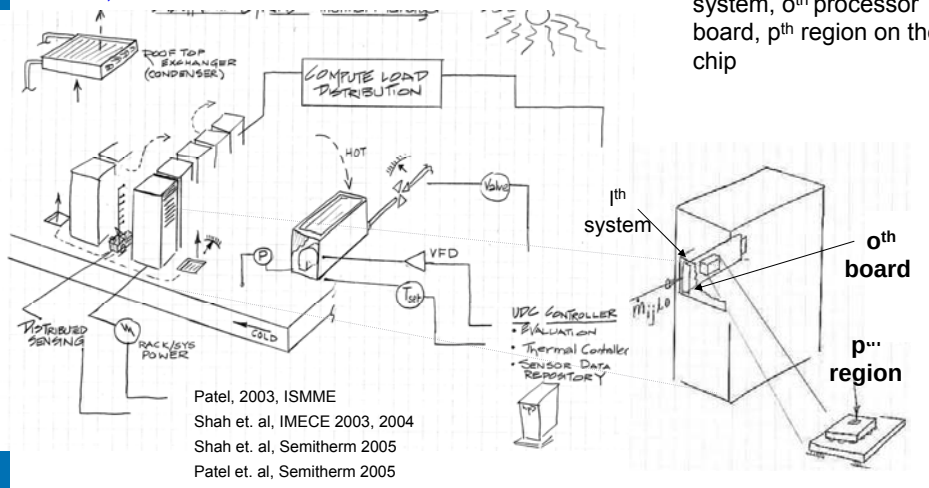
Dynamic Workload Placement to a data center in k^{th} geographic region*

* Patel et. al, IMECE 2003

Phoenix, AZ at 20°C

New Delhi at 45°C

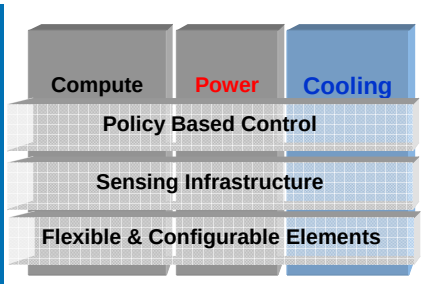
i^{th} row, j^{th} rack, l^{th} system, o^{th} processor board, p^{th} region on the chip



Patel, 2003, ISMME
Shah et. al, IMECE 2003, 2004
Shah et. al, Semitherm 2005
Patel et. al, Semitherm 2005

Summary

Smart Chip, System and Data Center



Rapid Commoditization of IT resources has resulted in great deal of focus on economics of energy

- "Burdened" cost of power and cooling exceeds depreciation of IT equipment
  - Patel & Shah, HPL TR, DC Cost Model

Minimizing destruction of Available Resources

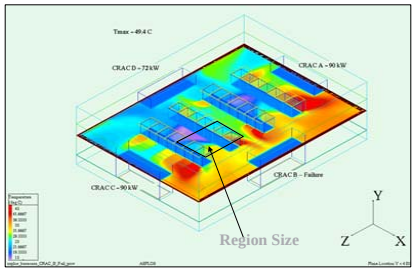
- Driven by Economics
- Pollution

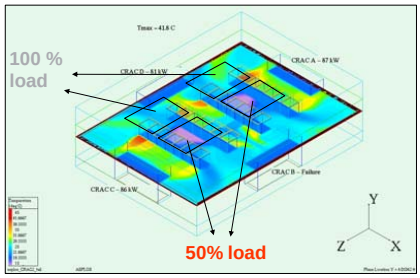
• Necessitates "smart" solutions as shown in this presentation to reduce the destruction of available resources by half – a quantifiable goal.

•Savings in energy consumption by cooling resources demonstrated at HP Labs smart data center

• Opportunity for a multi-disciplinary team in *mechanical engineering, materials engineering, electrical engineering and computer science*

I thank you for your time





"A control volume drawn around a planetary scale collection of computing devices, the confines of which maintain balanced cooling and compute loads, saves a world of energy"

Phoenix

Tokyo

Energy Aware Computing

New Delhi

meter scale

micro-mechanical

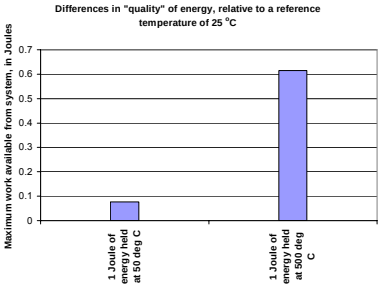
BACKUP

Developing Exergy-based Metric

HP- UCB, CITRIS Collaboration with Prof. Van Carey, Amip Shah UC Berkeley – CITRIS Collaboration)

What is exergy?

- **Not all energy is equally beneficial**
  - **Only energy which can be converted to work is useful**
- **Second law of thermodynamics: irreversible losses**
  - **More irreversibilities → less work can be extracted**
  - **Surroundings also impact availability**
- **Exergy measures "usefulness" of system by quantifying amount of energy available for work**



The same system at different conditions can have different levels of utility, a characteristic captured by the metric of exergy.

Reference:

J. Szargut, D. R. Morris, F. R. Steward. *Exergy Analysis of Thermal, Chemical and Metallurgical Processes*, Hemisphere Publishing Corporation, New York, NY (1976).

Global Power Use

reducing energy consumption - targets

- 6 billion devices at 50 to 100 W each, mostly network attached to global collection of data centers e.g. thin clients
  - 75W, 0.5 TW
  - Energy Aware Computing based on utilization & passive cooling solutions can reduce power consumption by 70%
- Specialized high performance desktops: 100 million
  - Future high performance chip cooling solution: 30 W
  - Active cooling solution without scale down: 3 GW
  - Scale down based on utilization: **Target 0.5 GW**
- Global distribution of data centers: US ~ 1500; rest of world: ~1500 [Future distribution can shift with global economy]
  - Power for hardware and cooling equipment in US: 3 GW
  - Power for hardware and cooling equipment – rest of world: 3 GW
  - Scale down based on utilization: **Target 0.5 GW***
  - ***** central nature of resources, and ROI, makes this the low hanging fruit**