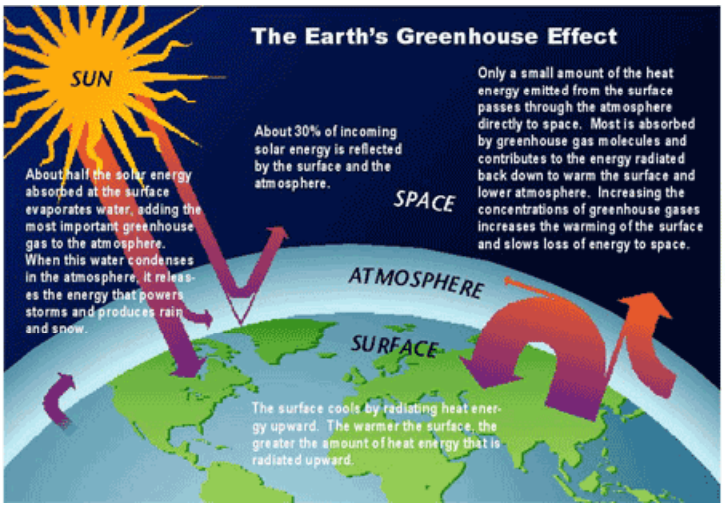
Energy-Efficient Computing

Massoud Pedram

University of Southern California
Dept. of Electrical Engineering
July 25, 2009

ACM SIGDA DA Summer School

Global Warming



Source: U.S. Global Change Research Program

2

M. Pedram, USC

Causes and Effects

Power plants, cattle and cars are some of the major contributors of greenhouse gases such as carbon dioxide and methane.

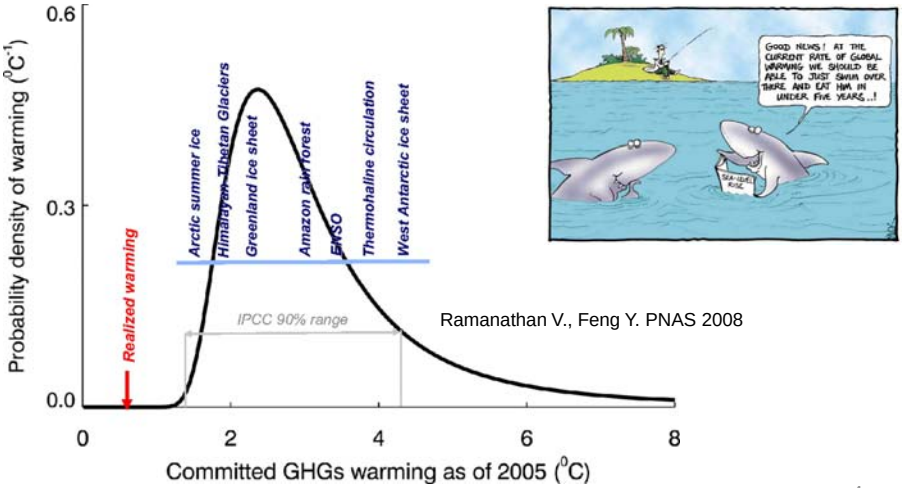
Some possible effects of global warming are the inundation of low-lying islands due to rising sea levels, increased frequency of severe storms and the retreat of glaciers and icecaps.

M. Pedram, USC

3

The Planet Is Already Committed to a Dangerous Level of Warming

Probability distribution for the committed warming by GHGs between 1750 and 2005



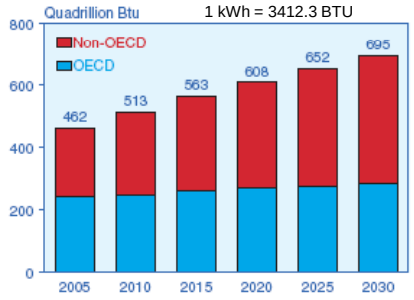
Ramanathan V., Feng Y. PNAS 2008

M. Pedram, USC

4

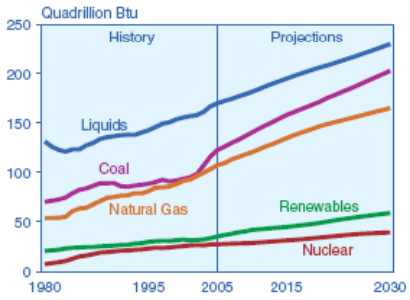
Energy Usage Worldwide

Figure 1. World Marketed Energy Consumption, 2005-2030



Sources: **2005:** Energy Information Administration (EIA), *International Energy Annual 2005* (June-October 2007), web site www.eia.doe.gov/iea. **Projections:** EIA, World Energy Projections Plus (2008).

Figure 2. World Marketed Energy Use by Fuel Type, 1980-2030



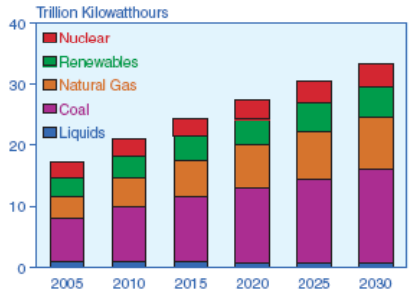
Sources: **2005:** Energy Information Administration (EIA), *International Energy Annual 2005* (June-October 2007), web site www.eia.doe.gov/iea. **Projections:** EIA, World Energy Projections Plus (2008).

World marketed energy consumption is projected to increase by 50 percent from 2005 to 2030. Total energy demand in the non-OECD countries increases by 85 percent, compared with an increase of 19 percent in the OECD countries.

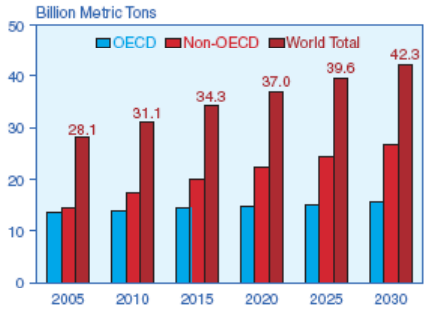
- Energy Information Administration / International Energy Outlook 2008

M. Pedram, USC

World Electricity Generation and Carbon Dioxide Emission



Sources: **2005:** Energy Information Administration (EIA), *International Energy Annual 2005* (June-October 2007), web site www.eia.doe.gov/iea. **Projections:** EIA, System for the Analysis of Global Energy Markets/Global Electricity Module (2008).



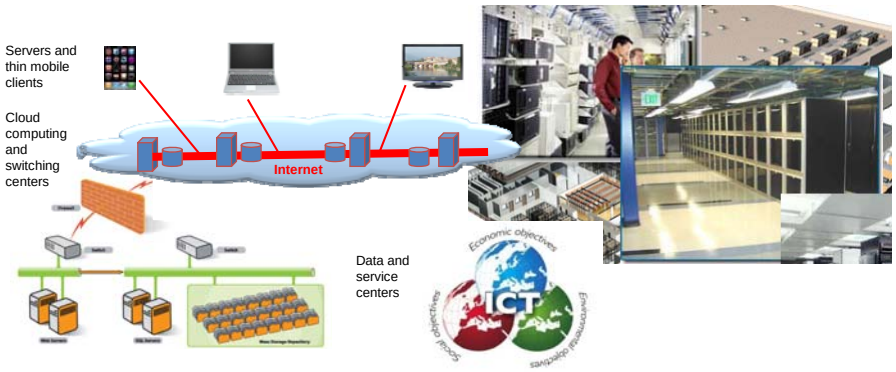
Sources: **2005:** Energy Information Administration (EIA), *International Energy Annual 2005* (June-October 2007), web site www.eia.doe.gov/iea. **Projections:** EIA, World Energy Projections Plus (2008).

Projections by the DoE's Energy Information Administration show that worldwide electric power demand will increase from the current level of about 2 Terawatts (TW) to 5 TW by 2050.

6

M. Pedram, USC

The ICT Ecosystem

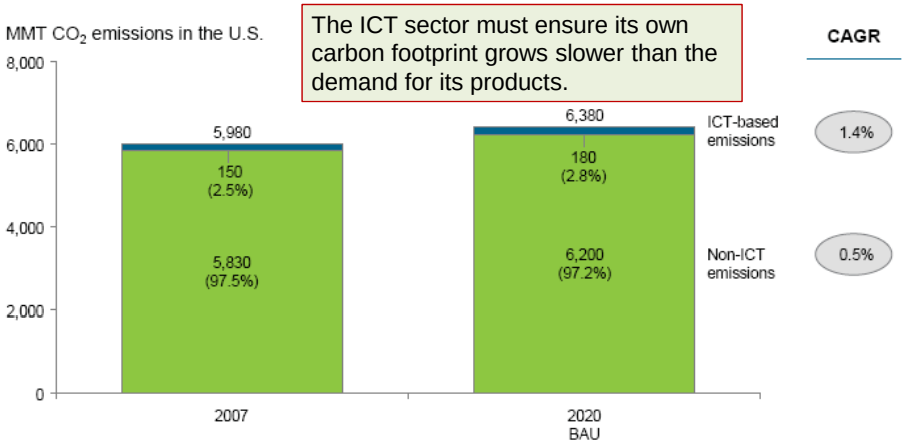


The ICT ecosystem encompasses the policies, strategies, processes, information, technologies, applications and stakeholders that together make up a technology environment for a country, government or an enterprise.

--Source: Harvard Law

M. Pedram, USC 7

ICT Industry's Carbon Emissions in the U.S.

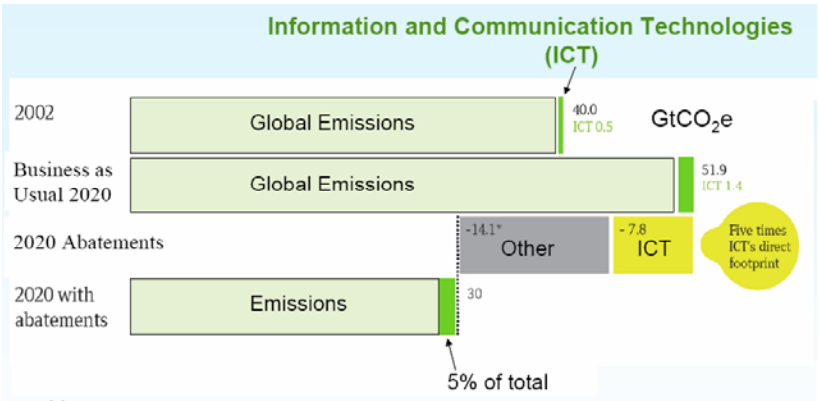


Sources: GeSI and the Climate Group, "Global Smart 2020 report", 2008; EIA Annual Energy Outlook 2008.

M. Pedram, USC 8

Target 2020

California's Global Warming Solutions Act of 2006
– Requires Reduction of GHG by 2020 to 1990 Levels
– 10% Reduction from 2008 Levels; 30% from BAU 2020 Levels
The European Union Requires Reduction of GHG by 2020 to 20% Below 1990 Levels (12/12/2008)



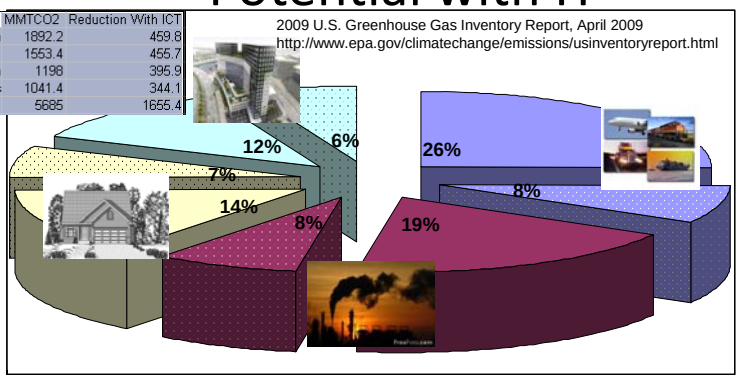
How the ICT Industry Can Reduce Its Own Carbon Footprint

- PC efficiency:
  - Substitution of energy-intensive desktops with laptops and thin clients, replacement of cathode ray tube monitors with LCD screens, and improvements in standby power management
- Telecommunication network and device efficiency
  - Reduction of standby power use of devices and greater adoption of network optimization processes
- Data center:
  - Server virtualization, adoption of best practices in heating and cooling and innovations in increased server efficiency improvements

Green House Gases: Reduction Potential with IT

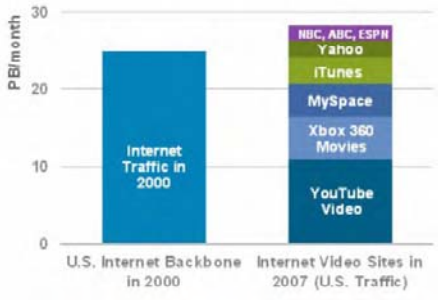
	MMTCO2	Reduction With ICT
Transporta	1892.2	459.8
Industrial	1553.4	455.7
Residentia	1198	395.9
Commerci	1041.4	344.1
Total	5685	1655.4

2009 U.S. Greenhouse Gas Inventory Report, April 2009
http://www.epa.gov/climatechange/emissions/usinventoryreport.html

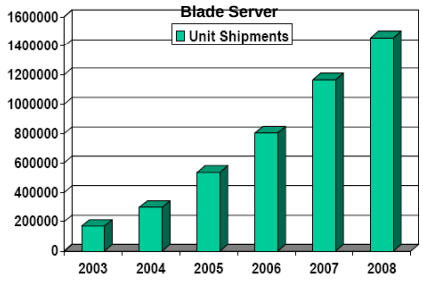


- Rising energy consumption and growing concern about global warming
  - World wide GHG emissions in 2006: 27,225 MMT of CO_2 Equivalent
- Major contributors: Transport, Industrial, Residential, Commercial
- Improved efficiency with Information Technology (IT) usage
  - 29% expected reduction in GHG
  - Equal to gross energy and fuel savings of $315 billion dollars
- Expect even wider utilization of IT in various sectors it is more energy-efficient

Internet Growth is Driving Data Center Usage



Source: ADC

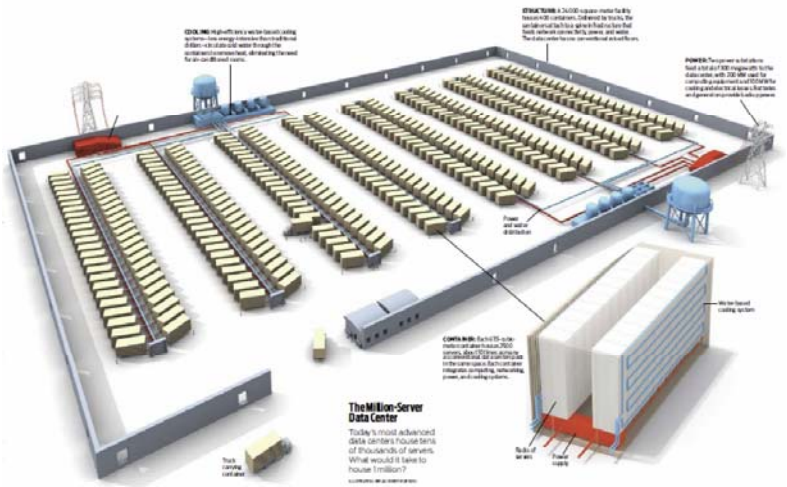


Increase from 500K to 1.45mn in 3 years

Source: Gartner

Corporate data growing 50 fold in three years.
—2007 Computerworld

Emergence of Mega Data Centers



Microsoft's Chicago Datacenter: Entire first floor is full of containers; each container houses 1,000 to 2,000 systems; 150 - 220 containers on the first floor.

Data Center Energy Efficiency Emerging as a Strategic Opportunity

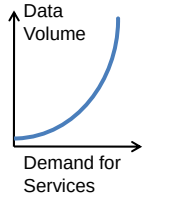
In 2006, datacenters used 1.5% (60 billion kW-hr/year) of all the electricity produced in the US … if nothing significant is done about the situation, this consumption will rise to 2.9% by 2011.
—"Report to Congress on Server and Data Center Energy Efficiency," EPA 2007

Rack power levels exceeding 40 kW/rack while datacenter power densities approaching and exceeding 500 W/sq ft.
—Tahir Cader, Power & Cooling Strategist, HP, April 2009

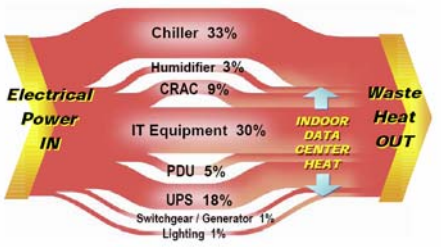
42% of datacenter owners said they would exceed power capacity within 12 to 24 months without expansion. 39% said they would exceed cooling capacity in 12 to 24 months.
—Infoworld, March 26, 2008

Datacenter construction costs projected to exceed $1,000/sq ft or $40,000/rack. —IDC's Datacenter Trends Survey, 2007

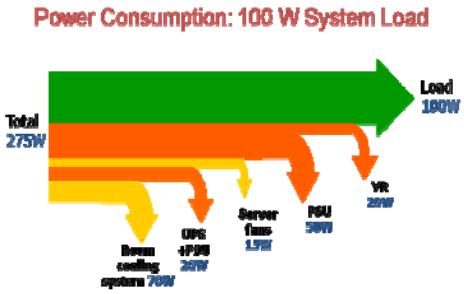
By 2010, the incremental demand for datacenter energy will equate to 10 new power plants. The new paradigm in datacenter metrics is shifting from square footage to capacity in megawatts.
—Processor.Com, February 20, 2009



Data Center Power Distribution



Source: IBM

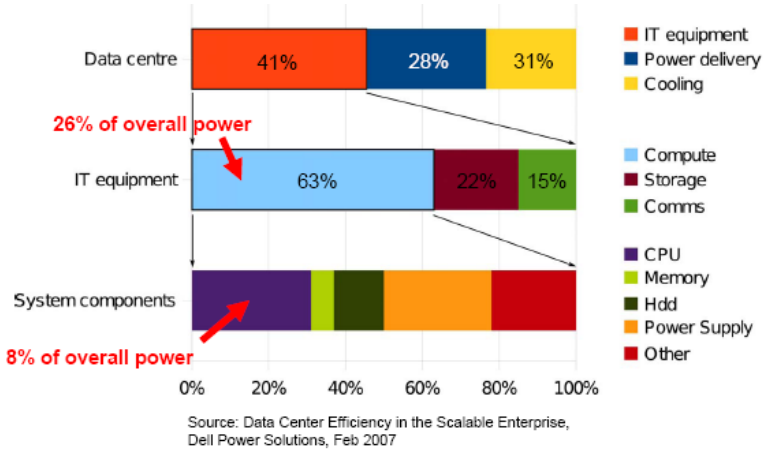


Source: Intel

More than 30 millions physical servers currently installed: growing 4X over the next 10 years.
For every $1 spent on server hardware, 50 cents is spent on power distribution and cooling.

— IDC, May 2008

Where is the Power Consumed



Source: Data Center Efficiency in the Scalable Enterprise, Dell Power Solutions, Feb 2007

Peak Power Demands of Data Centers

- Apart from the total energy consumption, another critical component is the peak power; the peak load on the power grid from data centers is currently estimated to be approximately 7 gigawatts (GW), equivalent to the output of about 15 baseload power plants.
  - This load is increasing as shipments of high-end servers used in data centers (e.g., blade servers) are increasing at a 20-30 percent CAGR.
  - If current trends continue, power demands are expected to rise to 12 GW by 2011.
  - Indeed, according to a 2008 Gartner report, 50 percent of data centers will soon have insufficient power and cooling capacity to meet the demands of high-density equipment.

Green IT: Reducing Oil Dependence

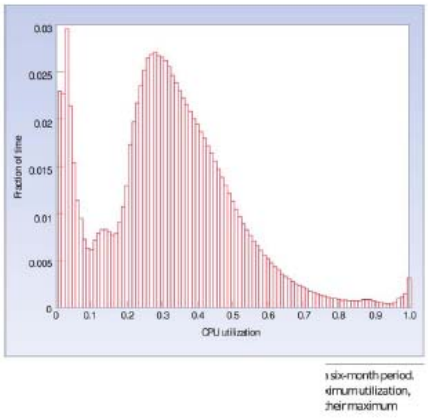
- US imports about 3.5million barrels of crude oil per day from the Middle East and Venezuela
  - Energy content of a barrel of oil is ~1,700 kWh
- As of 2006, the electricity use attributable to nation's servers and data centers is estimated at 61B kWh
  - This is projected to double by 2011 and triple by 2015
- Assume a moderate reduction of 25% in energy consumption of the U.S. data centers
  - Reduces U.S. foreign oil imports by 98,000 barrels a day!!
- With wider adoption of green ICT, even higher reduction of oil imports will be achieved
  - 1kWh energy consumed in a data center can help eliminate x >1 (e.g., 5 to 10) kWh of energy in other sectors
  - Although the carbon emission due to every kWh of electrical energy consumed in the U.S. varies depending on the type of power generator used to supply power into the power grid, an average conversion rate of 0.433 kg CO_2 emission per kWh of electrical energy may be assumed

Approach to Energy Efficiency of ICT Solutions

- Need a multi-pronged approach to energy efficiency which is cross-tier, cross-area, and cross-discipline
  - Stake holders: users, industry, academics, government
  - Environmental constraints and socio-economic drivers
  - Policies and treaties
  - Technologies and applications
    - Hardware and server architectures
    - Storage
    - Networks and networking
    - Software and middleware
  - Physical infrastructure
    - Power generation (including renewable sources), The Grid – power distribution and delivery, demand shaping
    - Electromechanical, Electrochemical, lighting and Cooling/AC

Server Utilization

"The Case for Energy-Proportional Computing," Luiz André Barroso, Urs Hölzle, *IEEE Computer* December 2007

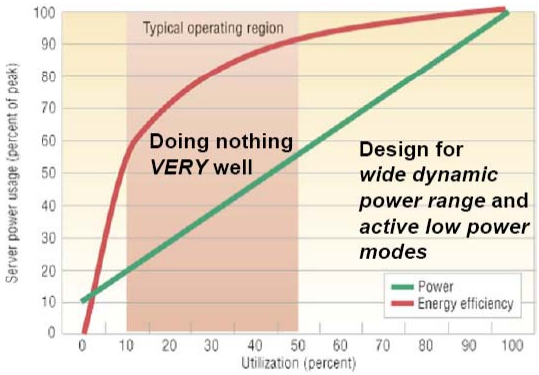


six-month period.
imum utilization,
their maximum

Servers are rarely completely idle and seldom operate near their maximum utilization, instead operating most of the time at between 10 and 50 percent of their maximum.

An Example Energy-Proportional Server

"The Case for Energy-Proportional Computing," Luiz André Barroso, Urs Hölzle, *IEEE Computer* December 2007



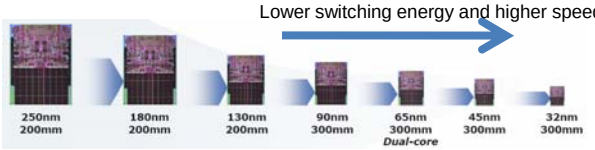
An energy-proportional server has a power efficiency of more than 80 percent of its peak value for utilizations of 30 percent and above, with efficiency remaining above 50 percent for utilization levels as low as 10 percent.

M. Pedram, USC

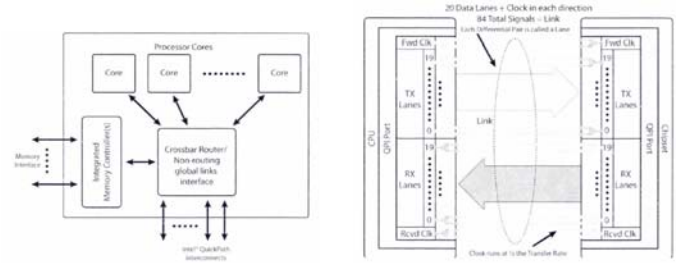
21

CMOS Process and Processor Design

- Moore's Law scaling (new CMOS process, the tick)



- Improvements in processor design and architectures (the tock)



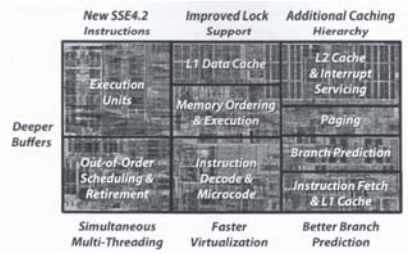
Intel Quickpath Interconnect: differential current-mode signaling; Each link is composed of 20 lanes per direction capable of up to 26.5 GB/s.

M. Pedram, USC

22

Micro-architecture and Server Architecture

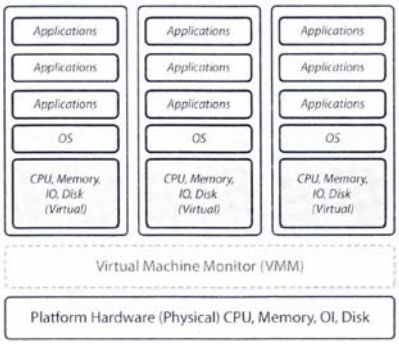
- Hyper threading (SMT) with appropriate Cache Hierarchy
- Integrated memory controller (support multiple channels of DDR3)
- PCI Express 3.0 (8Gbit/s per lane - will support 40 Gigabit Ethernet)
- On-chip(microcontroller-based) power management
- Selective speed boosting



Nahalem New Micro-architectural features

Virtualization

- Hardware assists for virtualization
  - Higher privilege ring for the hypervisor
  - Handoffs between the hypervisor and guest OS supported in hardware
  - Processor state info is retained for the hypervisor and for each guest OS in dedicated address space
  - Extended page tables
  - Virtual processor ID to avoid flushes on VM transitions

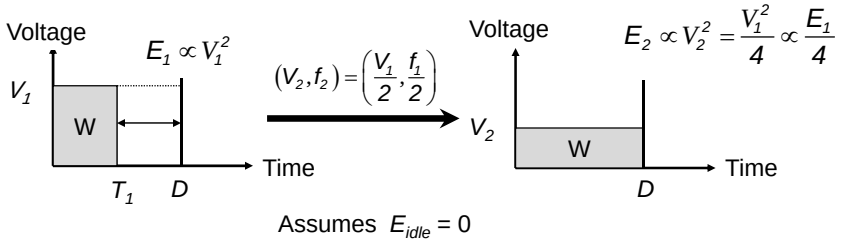


Dynamic Power Minimization Techniques

- Dynamic voltage and frequency scaling (DVFS)
- Multiple Voltage Islands (MVI)
- Dynamic Power Management (DPM)
- Green Data Centers

Dynamic Voltage and Frequency Scaling (DVFS)

- Energy, *E*, required to run a task during T: $E = P \cdot T \propto V^2$
- Example: a task with workload *W* should be completed by a deadline, *D*

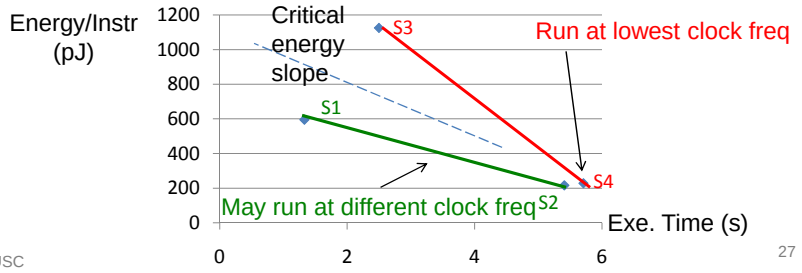


Assumes $E_{idle} = 0$

DVFS and IPC Value

- 1B instructions – Use a *critical energy slope* of 150pJ per instruction per sec

	IPC Value	Core clock frequency	MIPS	Core power	Execution Time	Energy per instruction
S1	High	600MHz	750	450mW	1.33s	595.5pJ
S2	High	180MHz	185	40mW	5.41s	216.4pJ
S3	Low	600MHz	400	450mW	2.5s	1,125pJ
S4	Low	180MHz	175	40mW	5.71s	228.4pJ

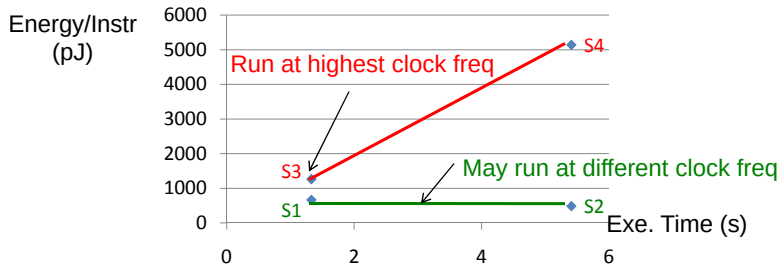


M. Pedram, USC

27

DVFS and Total System Power

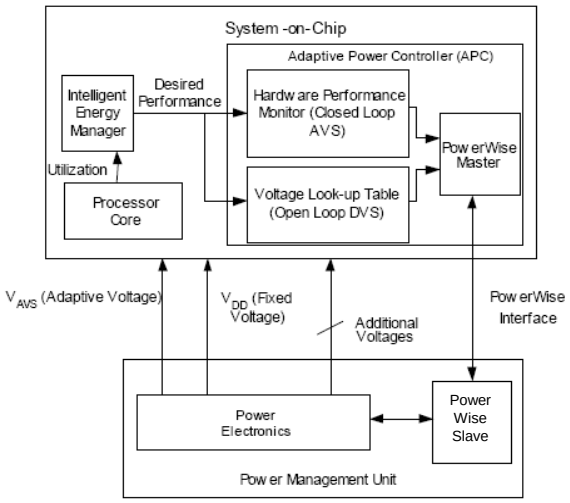
	Non-core power	Core clock frequency	MIPS (high IPC case)	Core power	Execution Time	Energy per instruction
S1	50mW	600MHz	750	450mW	1.33s	665pJ
S2	50mW	180MHz	185	40mW	5.41s	486.9pJ
S3	500mW	600MHz	750	450mW	1.33s	1,263.5pJ
S4	500mW	180MHz	185	40mW	5.41s	5139.5pJ



M. Pedram, USC

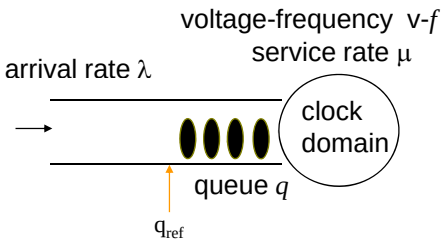
28

Open-Loop vs. Closed-Loop DVS

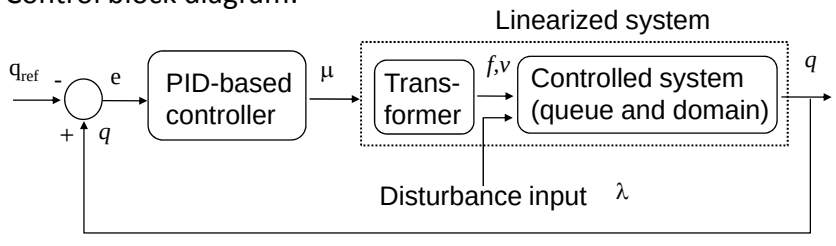


Adaptive Voltage-Frequency Scaling: A Linear Controller Design

- PID controller
  - Proportional gain (K_P)
  - Integral gain (K_I)
  - Derivative gain (K_D)



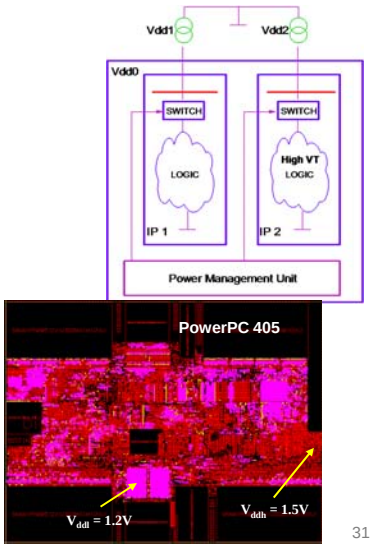
Control block diagram:



Source; Clark, HPCA-05

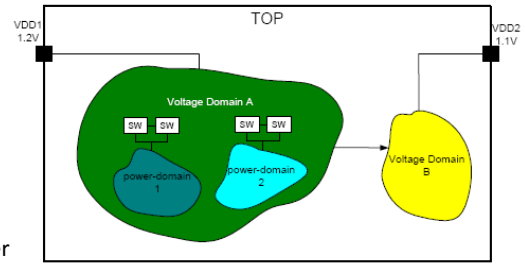
Multiple Voltage Islands

- Run power-isolated blocks at different V_{DD} levels
  - Can use mix of Low and High V_{th} devices to balance switching speed and leakage
  - Switch off inactive blocks to reduce leakage power dissipation
  - Requires library with isolation cells and level shifters
  - Requires library characterization data for multiple supply voltages
- An effective low power solution, but
  - Realizing its full potential requires integration of process technology, transistors, cell libraries, design methodology and tools



Voltage Islands and Power Domains

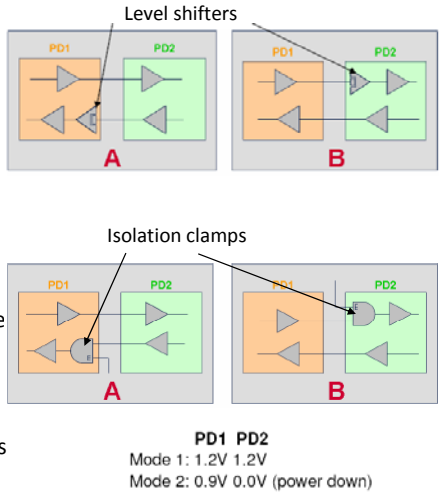
- **Voltage islands**: Areas (logic and/or memory) on a chip which are supplied through separate, dedicated power feeds
  - Typically, an 'always-on' voltage island is needed to implement critical functions such as real time clock, register file, PC counter, etc.
- **Power domains**: Areas within an island which are fed by the same V_{DD} source, but are independently controlled with an intra-island header switches



Example: IBM Blue Logic® Cu-08 voltage islands
L_{drawn} = 70 nm
Up to 72-million wireable gates
Power supply: 1.0 V with 1.2-V option
Power dissipation: 0.006 µW/MHz/ gate
Gate delays: 21 picoseconds (2-input NAND gate)
Eight levels of copper for global routing

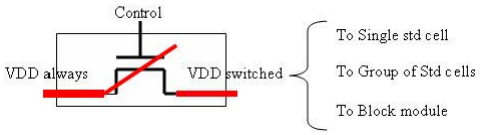
Level Shifters and Isolation Clamps

- Cannot directly connect V_{DDL} and V_{DDH} cells
  - Output of a V_{DDL} gate cannot go higher than V_{DDL}
  - When connected to a V_{DDH} gate, the PMOS transistor cannot be completely cut-off → Static Current
- Cannot directly connect output of a powered down block
  - Can propagate unwanted data in the driven logic
  - Floating input will potentially generate short circuit current
  - Specific inactive state or reset values may be forced on inputs driven by the power down logic



On-chip Power Switches

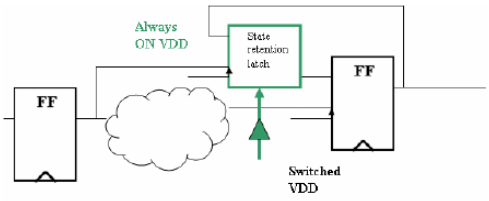
- Voltage islands are turned ON/OFF with on-chip switches



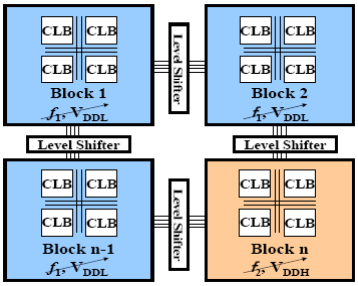
- Constraints and requirements
  - Power distribution and floorplanning are more complex tasks
  - Power switches need proper sizing: Balance switching current carrying capacity vs. layout area and leakage currents
    - Static IR drop analysis is necessary to verify sizing
  - Sequencing of the control signals influences the length of wake-up time and the amount of inrush current
    - Transient analysis is required to assess the impact of a switching block on its surrounding blocks

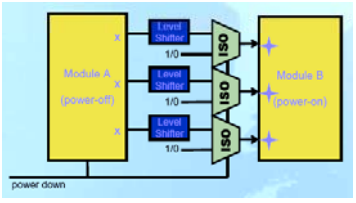
Data Retention

- Many possible variants
  - Integrated cell, shadow register with or without save and restore
- Constraints and requirements
  - A logic block with retention requires multiple voltage supply levels
  - Retention may require an always-on buffer tree for the control signals
  - Connection of supplies is error prone and time consuming
    - Difficult to rely on a global connection by simply doing pattern matching on pin or instance names



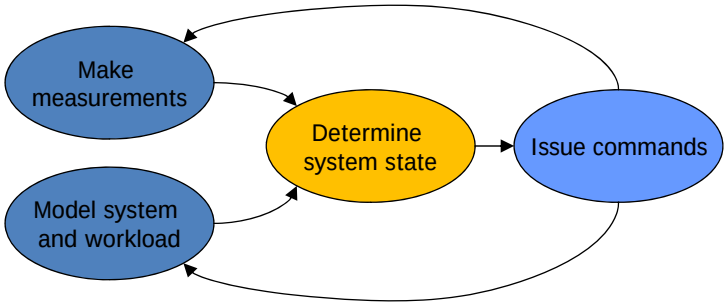
Combining DVS and MVI





- A logic-swing level shifter is a key circuit component in the DVS or MVI implementation because even for a chip-level DVS system, I/O blocks are operated under fixed V_{DD} and consequently level shifters are required between core circuits and I/O circuits
- When the DVS technique is applied on a "block level" to a MVI system as shown above, each block on the chip could be operated under either high V_{DD} (V_{DDH}) or low V_{DD} (V_{DDL}). Thus, voltage level shifters are needed among blocks to avoid static short circuit current at a receiver side

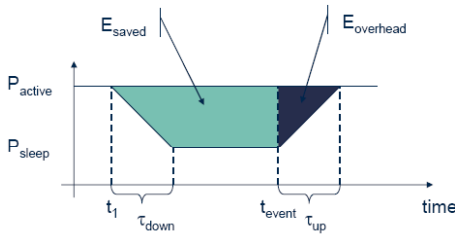
Generic Flowchart of a System-Level Power Manager



- Use hardware performance counters to gauge processor activity
  - Analyze program phases and adapt processor state accordingly
  - Recognize power/thermal hotspots and take (possibly preventive) action

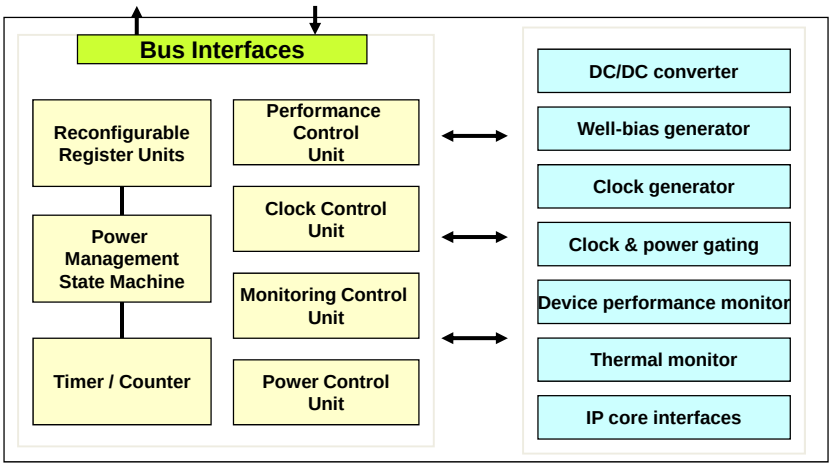
Switching Between Modes

- Simplest idea: Greedily switch to lower mode whenever possible
- Problem: Time and power consumption required to reach higher modes are not negligible
  - Introduces overhead
  - Switching only pays off if $E_{saved} > E_{overhead}$



- Example wakeup strategies:
  - Pre-scheduled wakeup
  - Event-triggered wakeup from sleep mode

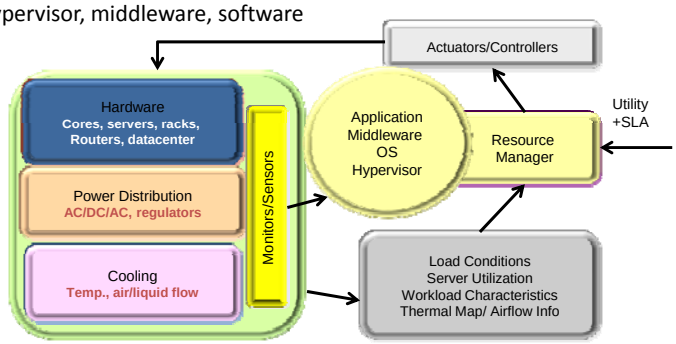
Typical Functions of an On-chip Power Management Unit



M. Pedram, USC 39

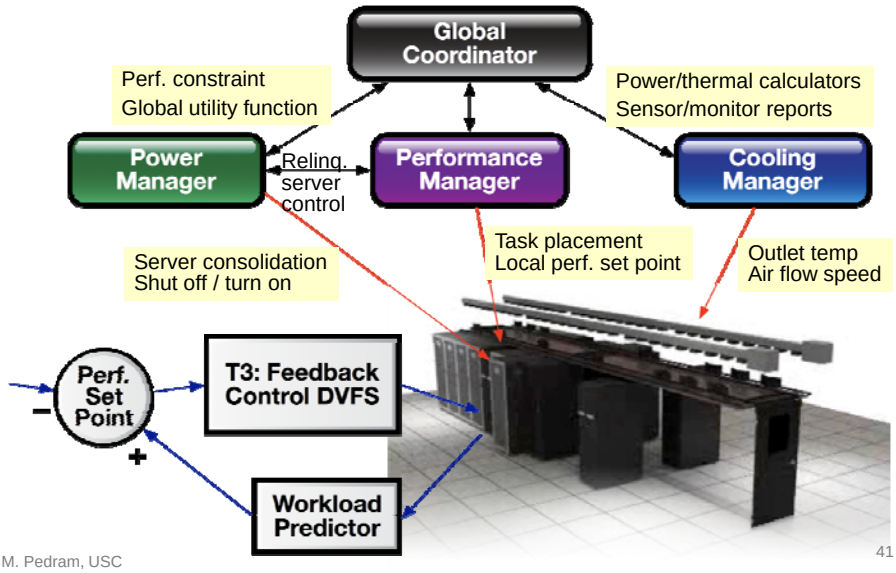
Adaptive Control in Datacenters

- Power management
  - Hierarchical, adaptive (learning-based), rigorous (control theory, game theory), and robust (models and manages uncertainty)
  - Receives runtime information from various instrumentation and holistically manages all datacenter resources
  - Utilizes various power/energy optimization levers provided in hardware, hypervisor, middleware, software



M. Pedram, USC 40

Multi-Agent, Multi-Tiered Power & Thermal Control

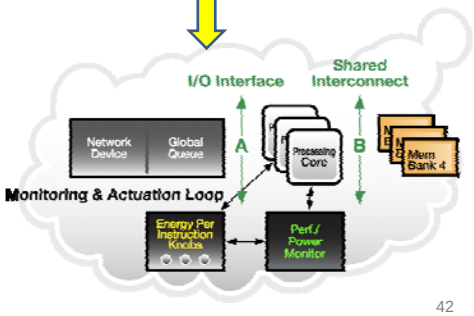


M. Pedram, USC 41

Energy-Aware Application Design

Start
Initialization
Read Data
PragmaStart(Mem-bound & Perf-Critical)
Until ? do Kernel1
PragmaEnd
Data transformation
Until ? do Kernel2
….
Prepare Output
End

Multiple Abstraction & Control Layers

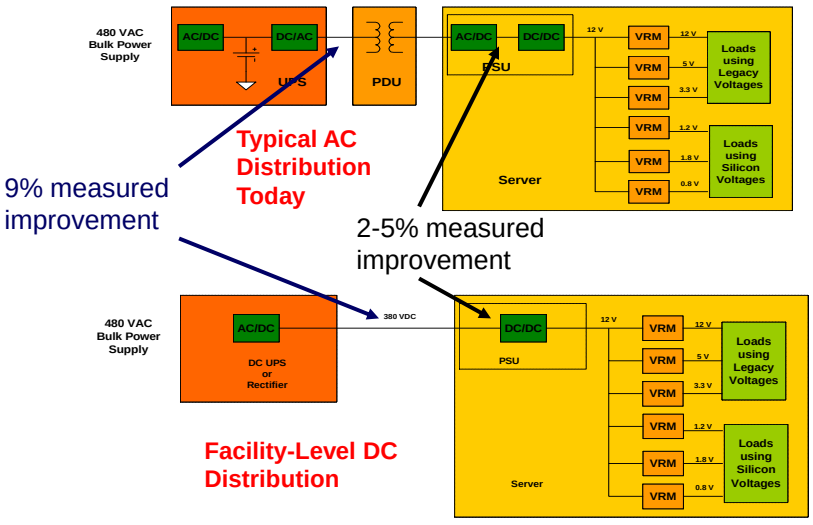


M. Pedram, USC 42

Physical Infrastructure

- But it's not the CAPEX – It is about OPEX (power and cooling)
  - By 2009, about half the world's data centers will be functionally obsolete due to insufficient power and cooling capacity to meet the demands of high-density equipment.
- New Cooling methods explored
  - In-server, in-rack and in-row cooling
  - Liquid cooling at the rack is 75%-95% more efficient than air cooling by a CRAC - Removes 100,000 Btu/hr of heat (~30 kW).
  - Cooling systems improvements call for improved airflow, optimization of temperature and humidity set points, and upgrades to water cooled chillers with variable speed fans and pumps.
- The price of power and cooling matters more and the geography matters less

Data Center Power Delivery System



It is About More than Watts …

- The real metric is services per joule per dollar
- Many interrelated ideas:
  - Application efficiency and energy management software
  - Workload variations and service level agreements
  - Micro-architecture and system design
  - Storage and network bandwidth and cost
  - Power availability and cost, power distribution and conversion efficiency
  - Packaging and cooling, component failure
  - Cost of people and money
  - Metrics that reward and punish