Network Swapping

Emanuele Lattanzi, Andrea Acquaviva and Alessandro Bogliolo

STI – University of Urbino, ITALY

Outline

- Motivations
- HW and SW support for swapping under Linux OS
  - Local devices (CF, μHD)
  - Network devices (remote storage resources)
- Characterization of swapping costs
  - Metrics definition
  - Experimental setup
  - Inherent swapping cost
  - DPM effectiveness evaluation
- Mobility management for network swapping
  - Proposed infrastructure
  - Implementation
  - Testing

Why Network Swapping?

- Storage memory is limited in palmtop PCs
  - Size, weight, power
  - Data storage memory and main memory frequently are implemented on the same physical space (DRAM)
  - Virtual memory is not useful
- Swapping needs in handheld devices
  - Limited application memory footprint
  - Only a few processes can run at a time
  - Context switch needs virtual memory to save process data and status when a process is not running
- Unlimited swapping space can be found on remote devices

UNIVERSITY OF URBINO - ITALY

ISTITUTO DI SCIENZE E TECNOLOGIE DELL'INFORMAZIONE

HW and SW Swapping Support

Linux Swapping

- Linux page-based swapping
  - Number of free pages falls below a threshold
  - A memory request cannot be satisfied
  - Approximate LRU global page replacement
  - Page size = 4Kbytes
- Heterogeneous support for swapping
  - Local
  - Remote using network support (NFS or NBD)

Swap Devices

- Compact Flash
  - Best performance at lowest energy
  - 32MB -1GB
- Microdrives
  - Fast evolution but still long seek time
  - OS software caches in main memory (trade-off with page requests) to perform bursty accesses
  - 340MB – 10 GB
- WNICs
  - Allows remote storage space
  - Use with NBD or NFS by mean of server-client demons

NFS vs. NBD

- Network File System
  - **UDP** – based
  - Client-server communication handled by **RPCs** (Remote Procedure Calls)
  - Swap area is a **remote file**
- Network Block Device
  - **TCP** – based (*larger protocol overhead*)
  - **No RPCs** (*smaller software overhead*)
  - Offers local block device interface to the OS
  - Swap area is a **remote device**

Characterization of Swapping Costs

Metrics definitions

- **Per Page Swap Time (PPST)**
  - Time spent metric
- **Per Page Swap Energy (PPSE)**
  - Energy consumption metric

Per Page Swap Time (PPST)

- **PPST** is the minimum time required to swap a single page from a given device
- Page swap involves
  - Transmission of page request
  - Waiting time due to latency of storage device
  - Reception of page
  - Possible write-back of swapped-out page
- **PPST** averages all these contributions

Per Page Swap Energy (PPSE)

- **PPSE** is the minimum energy required to swap a single page from a given device
- Page swap involves
  - Transmission of page request
  - Waiting time due to latency of storage device
  - Reception of page
  - Possible write-back of swapped-out page
- **PPSE** averages all these contributions

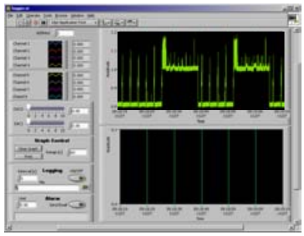
UNIVERSITY OF URBINO - ITALY
ISTITUTO DI SCIENZE E TECNOLOGIE DELL'INFORMAZIONE

Experimental Setup

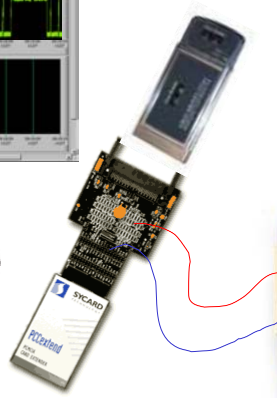
- Palmtop PC
  - IPAQ 3600 running Linux 2.4
  - 32MB main memory
- PCMCIA devices
  - CF = 64MB, μHD = 340MB, WNICs
- Power measurements
  - Sycard card extender
  - DAQ + Labview

© Emanuele Lattanzi SFM-05:Moby

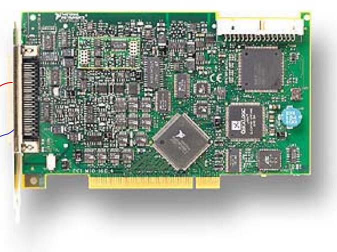
Current Measurement



- Current waveform acquisition software → LabView



- AD acquisition card
  - 200 ks/sec
  - 12 bit/sample



Characterization Benchmark

- Allocate and initialize a large matrix (bigger than main memory)
- Read by column to force page faults
- The swap device is always busy servicing swap requests

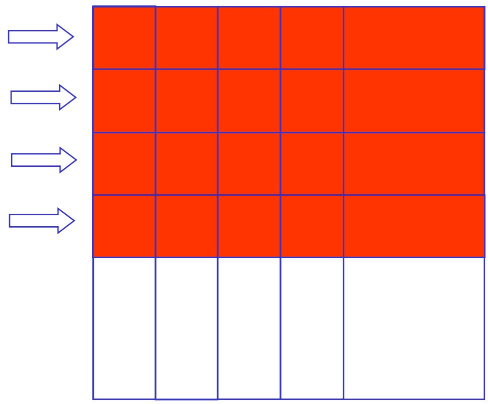
```
double A[ROW][COL];
initialize(A,ROW,COL);
t0 = time();
read_by_column(A,ROW,COL);
t1 = time();
```

Pseudocode of the benchmark used to characterize swap devices

- Replace read with write to characterize write-back

Page Swap Cost

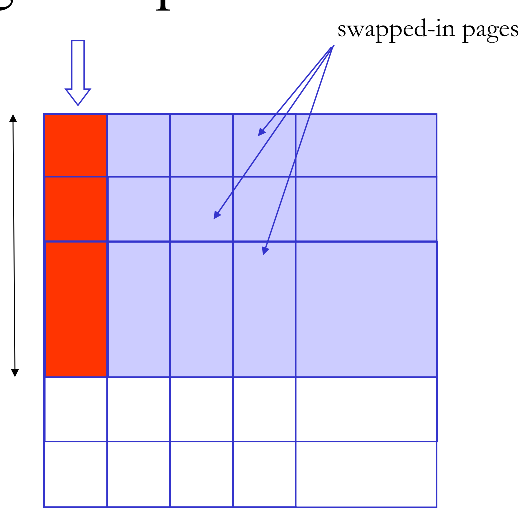
- Allocation by row



UNIVERSITY OF URBINO - ITALY
ISTITUTO DI SCIENZE E TECNOLOGIE DELL'INFORMAZIONE

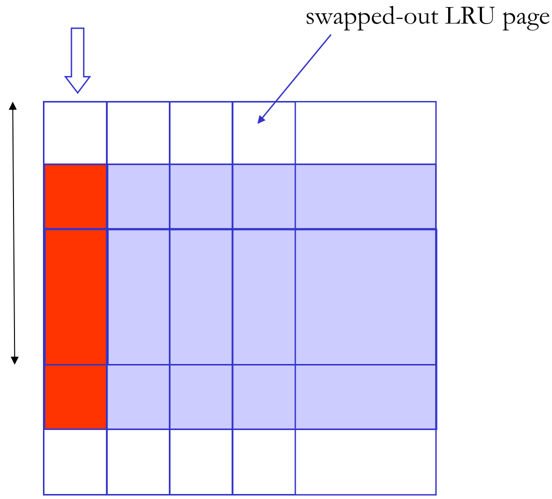
Page Swap Cost

- Access by column



Page Swap Cost

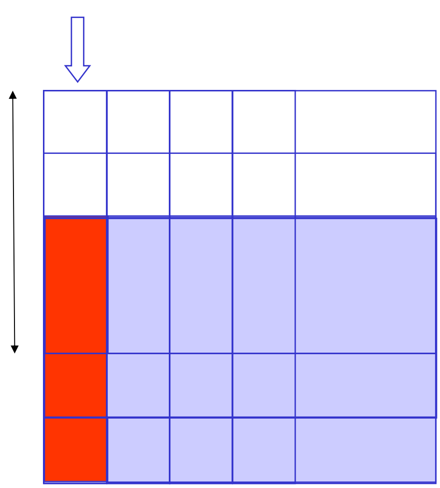
- Access by column



UNIVERSITY OF URBINO - ITALY
ISTITUTO DI SCIENZE E TECNOLOGIE DELL'INFORMAZIONE

Page Swap Cost

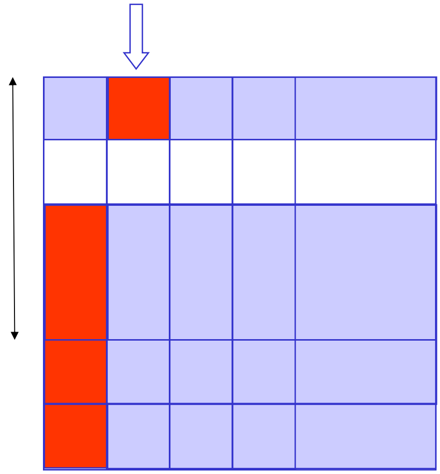
- Access by column



© Emanuele Lattanzi SFM-05:Moby

Page Swap Cost

- Each access to the matrix causes a page fault



Characterization Results

POWER MANAGEMENT DISABLED

Swap device		Read-only			Write-back		
Type	Mode	Time [ms]	Energy [mJ]	Power [mW]	Time [ms]	Energy [mJ]	Power [mW]
CF	local			49			49
HD	local			637			637
NIC_{CISCO}	NBD			848			735
NIC_{CISCO}	NFS			720			720
NIC_{COMPAQ}	NBD			578			573
NIC_{COMPAQ}	NFS			524			485
		PPST	PPSE		PPST	PPSE	

- Write back doubles the cost (as expected)
- Local devices are more efficient than WNICs
  - CF energy-per-page 10 times lower than other devices
- NBD provides greater performance than NFS

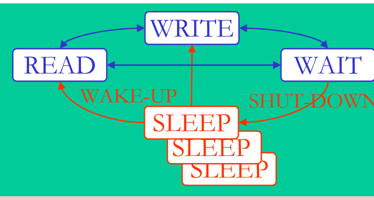
Characterization of DPM Support

- DPM Support Efficiency
  - Device energy states
  - Transition energy and time

Local Devices

CF, HD are timeout based

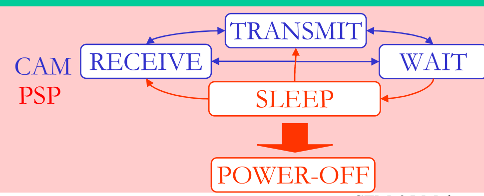
Timeout can be dynamically tuned based on workload



Remote Devices

MAC-level DPM: PSP

POWER-OFF: need software support



OS-Level DPM

- WNICs radio can be shut-off via software
- We modified NBD driver to issue shut-off system calls when no swap requests are performed
- Wake-up time state is large (re-association), but
  - No energy spent at all
  - Transition is controlled by NBD driver, which can exploit information about swapping activity to preemptively wake-up WNIC

Characterization Results

DYNAMIC POWER MANAGEMENT ENABLED

Device	State	Power [mW]	Timeout [ms]	WU-time [ms]	WU-power [mW]
CF	Read	107			
	Write	156			
	Wait	4.5			
HD	Read	946			
	Write	991			
	Wait	600			
	Sleep	24	2000	4500 ± 1980	1067
NIC_{CISCO}	Receive	755			
	Transmit	1136			
	Wait	525			
	Doze(PSP/PSPCAM)	113	0/850	14/14	400
	Power-Off	0	any	370	451
NIC_{LUCENT}	Receive	548			
	Transmit	798			
	Wait	407			
	Doze	38	100	1	800
	Power-Off	0	any	270	357

large wake-up delay
but OS controlled

DPM effectiveness evaluation

DPM effectiveness evaluation

- Characterization benchmark is unrealistic for two main reasons:
  - Computation time is usually non-negligible → swap requests are spaced in time
  - The total size of the data structure accessed by an application usually does not exceed the size of the main memory
- We use a case-study benchmark to evaluate the DPM effectiveness

Case Study - Benchmark

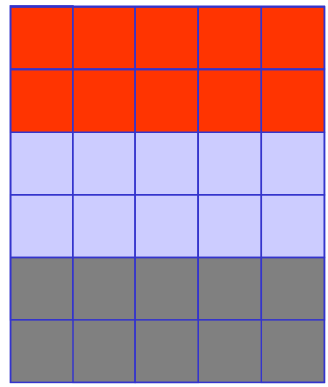
- A, B, C fits in memory, but dummy initialization swaps them out
- Product computation generates swap requests

```
double dummy[2048][2048], C[128][128];
double A[128][128], B[128][128];
initialize(A,128,128);
initialize(B,128,128);
initialize(C,128,128);
initialize(dummy,2048,2048); //swap out
t0 = time();
compute_product(A,B,C);
t1 = time();
```

Pseudo-code of the case study

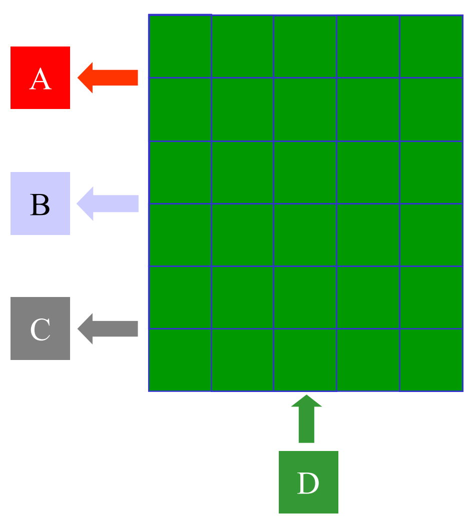
Benchmark

- A, B, C allocation and initialization



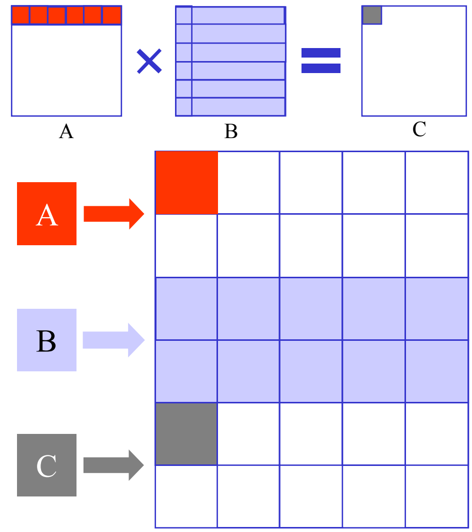
Benchmark

- A, B, C allocation
- Dummy allocation



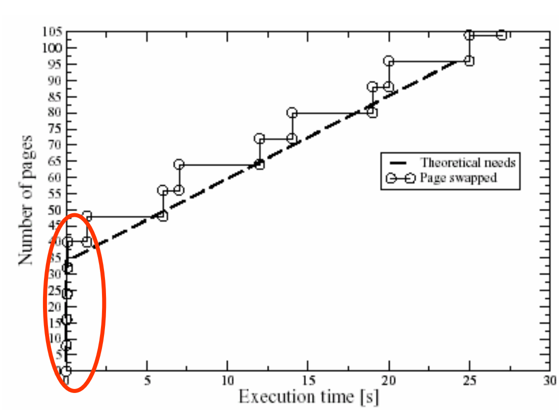
Benchmark

- A, B, C allocation
- Dummy allocation
- Product
  - First entry of C -> read first column of B
  - Upload of entire B (read by column)



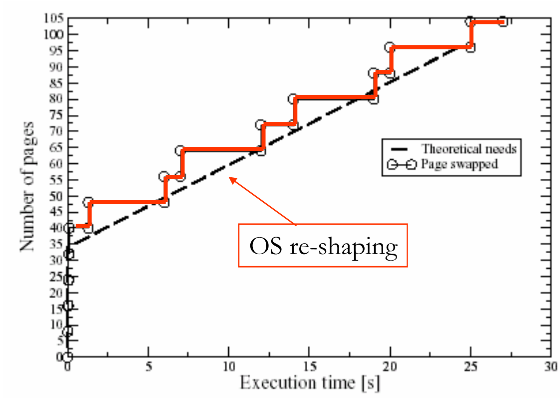
Page Requests

- First burst of request due to upload of B



Page Requests

- First burst of request due to upload of B
- OS requests 8 pages at a time



Dynamic Power Management

RESULTS

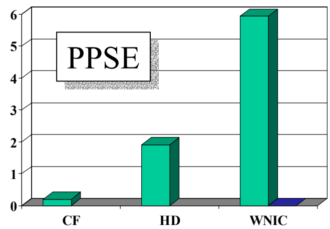
- HD spends more energy than WNICs (large idle power)
  - HD DPM is counterproductive in time and energy
- WNIC DPM saves 50-80% of swapping energy
  - 85-94% with power-off state

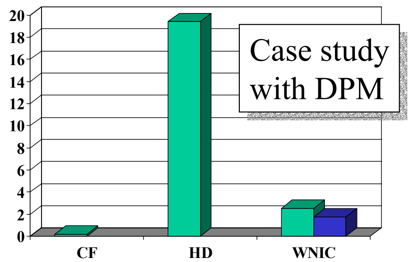
BENCHMARK EXECUTION

Device	Exec. time [s] Avg	Exec. time [s] Std	Energy [mJ] Avg	Energy [mJ] Std
RAM	25	0	-	-
CF	25.5	0.57	0.14	0.003
HD	*25.31*	-	*15.20*	-
(PM ON)	**37.75**	5.91	[illegible]	5.31
NIC_{CISCO}	26	-	16.51	0.01
(PSPCAM)	**30.67**	2.16	**10.59**	0.23
(PSP)	**43.33**	0.58	**8.05**	0.23
(OFF)	**28.75**	0.5	2.47	0.09
NIC_{LUCENT}	30.25	0.5	13.60	0.56
(PSP)	**30.0**	0	[illegible]	0.096
(OFF)	**30.0**	0	[illegible]	0.08

HD vs WNIC

Results Summary



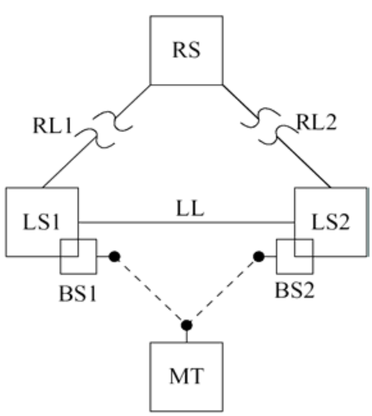


Mobility Management for Network Swapping

Mobility Management for Network Swapping

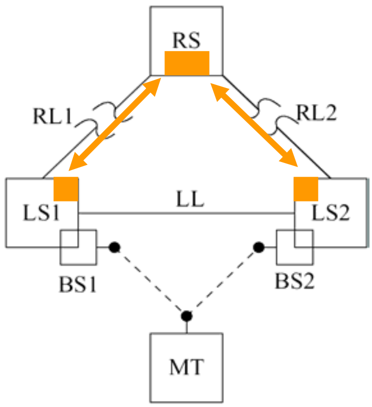
- Providing network virtual memory across a wireless link raises many issues
  - Performance and energy efficiency of WNICs
  - Performance of the network
  - Service management during terminal mobility
- We focus on keep the virtual memory pages local to the base station granting wireless network connectivity and QoS to the mobile terminal
- We need an infrastructure providing content migration during access points handoff events

Infrastructure overview



- Base stations BS1 and BS2 provide wireless network connectivity to mobile terminal (MT)
- Base stations are connected to the local server LS1 and LS2
- Local servers are connected by the wide-band link LL
- A remote server RS is connected to LS1 and LS2 through multi-hop remote links RL1 and RL2
- The mobile terminal is associated to BS1 and move toward the service are covered by the BS2

A possible solution: Remote swap area



- The swap area of the Mobile terminal is unique and made available by the RS
- LS1 and LS2 provide proximity cache to enhance performance and mobility.
- Management mechanisms are required to guaranty coherency

Handoff event

- An handoff event involves two phases:
  - De-association from an access point
  - Re-association to a new access point
- During de-association and re-association there is no connectivity between MT and BSs and the service is suspended

Write-back policy

- **Phase A**: The MT de-associates from BS1 and simultaneously LS1 start to write-back all dirty pages on the local cache to the RS
- **Phase B**: The MT re-associates from BS1 a new empty cache is instantiated on LS2. The new cache is empty and its efficiency is initially lower then that of the previous cache

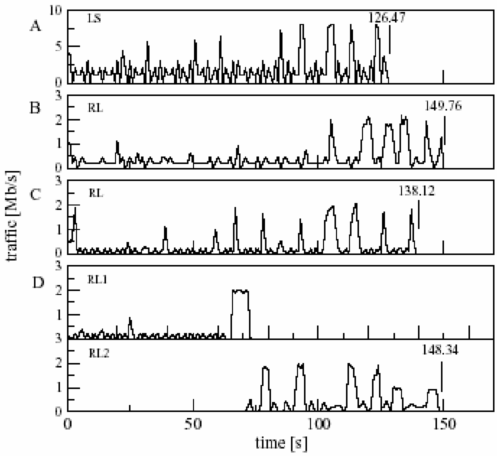
UNIVERSITY OF URBINO - ITALY
ISTITUTO DI SCIENZE E TECNOLOGIE DELL'INFORMAZIONE

Infrastructure implementation

- We use a remote swap server based on the NBD provided by Linux OS
- Particular issues about mobility management:
  - Prediction and control the 802.11b handoff mechanism to exploit its black-out time to perform swap service migration
  - Implementation of proximity caches and buffers
  - Dynamically switching from a NBD server to another

© Emanuele Lattanzi
SFM-05:Moby

Infrastructure tests



- **A**: swap area installed on a local server
- **B**: swap area installed on remote server
- **C**: remote swap area with a proxy cache local to the BS
- **D**: handoff event with write-back policy

Conclusions

- We made a comparative analysis of local and remote swap devices for palmtop PCs
- We demonstrate that network virtual memory for MT in a wireless scenario can be competitive both in term of power and performance
- We have presented an infrastructure that provides efficient support to network virtual memory for mobile terminal
- We have proposed a simple strategies to keep the virtual memory pages local to the base station granting wireless network connectivity to the mobile terminal