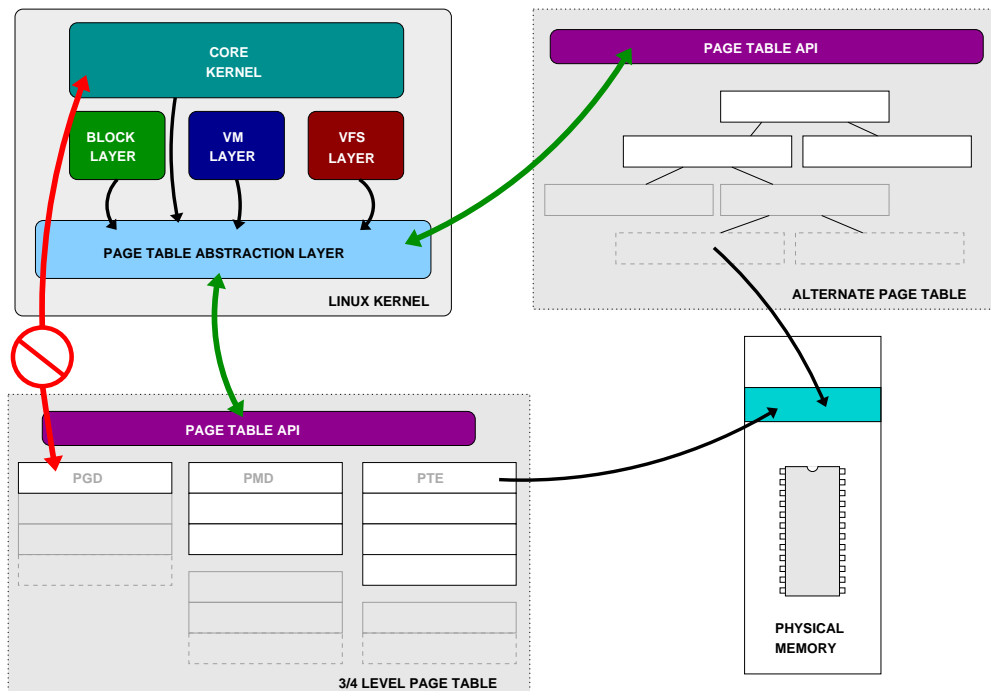


# Page Table Abstraction

- Most of the core kernel **assumes** a four level hierarchical page table.
- The kernel can be **abstracted** from the page table implementation.
- The kernel has been modified to only access page tables via a **defined interface**.
- Page table implementations implement the page table API.
- Page tables could be swappable on boot, or one day, per process depending on workload.



## Alternative Page Table Implementations

- Alternative page table formats can be tuned to architecture and workload.
  - Guarded Page Table
  - B-Tree Page Table
  - Hashed Page Table

## Research Questions

- How much overhead does abstraction introduce?
- Support for other architectures
  - POWER would fit very well
  - x86 assumes table shape
- Quantify benefits of different page tables.

# NFS Performance

## The Problem

- People report NFS is **Slow**
- Developers use **SpecSFS '97** for performance evaluation.
  - SpecSFS '97 is **almost ten years old** — there's no guarantee it reflects **current usage patterns**.

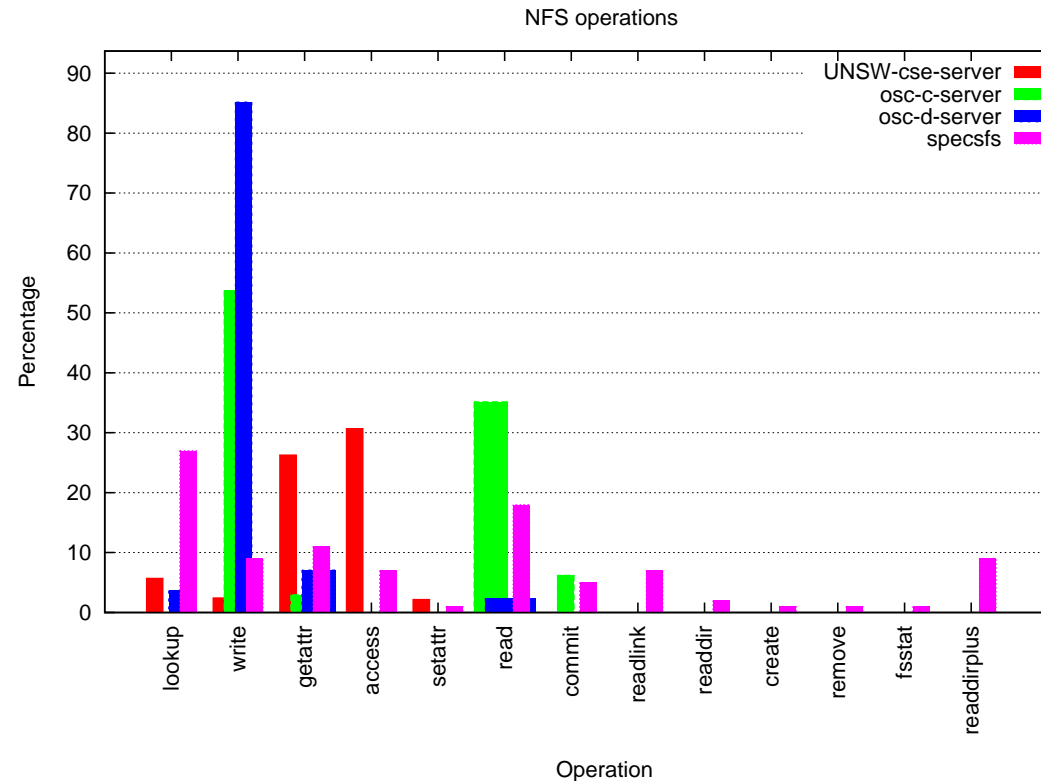
## Our Work

- Gather anonymised **traces** from as many NFS servers as we can.
- **Analyse** traces for patterns.
  - **Markov Matrices** for trace regeneration.
  - **Operation Probability Distributions** for feeding to SFS driver.
- (future) **Replay** traces at accelerated rate to find server bottlenecks.
- (future) **Optimise** server for observed patterns.

## Can you help?

- We need more traces — see <http://www.gelato.unsw.edu.au/IA64wiki/NFSBenchmarking/> for details
- Contact us if you have interesting work loads, and can supply anonymised traces.

## Workload Characteristics



Usage pattern for NFS version 3.0, on three different servers, and the pattern used for Spec SFS '97 version 3.0. Note how different the SFS pattern is from any observed pattern

# Community Involvement

## Projects

### IA64 Wiki

Collaborative documentation effort  
<http://www.gelato.unsw.edu.au/IA64wiki>

### Mailing Lists and Archives

Mailing Lists relating to Gelato interests  
<https://www.gelato.unsw.edu.au/mailman/listinfo/>

### Gelato Debian Repository

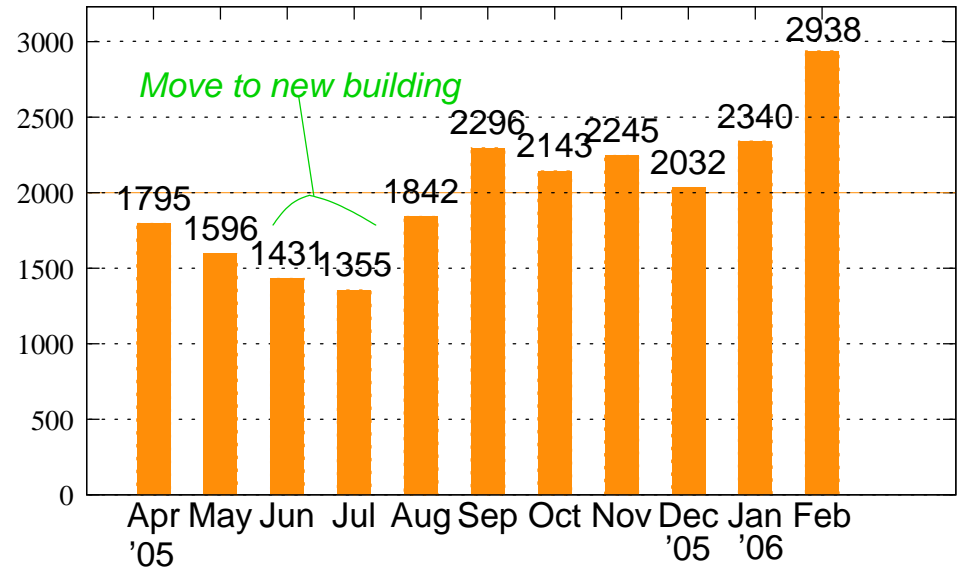
A central source of IA64 Linux software for Debian  
<http://www.gelato.unsw.edu.au/IA64wiki/GelatoDebianRepository>

### Kernel Autobuild

Checking kernel builds and patches daily (now via git)  
<http://www.gelato.unsw.edu.au/kerncomp>

## Website

Gelato@UNSW Web Average Daily Visits



## Recent Publications

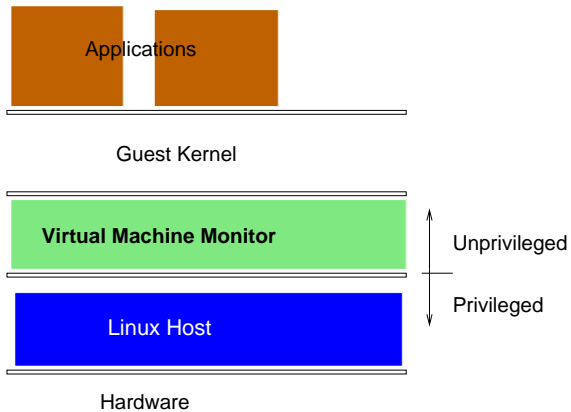
- Joshua LeVasseur, Volkmar Uhlig, Matthew Chapman, Peter Chubb, Ben Leslie and Gernot Heiser  
**Pre-virtualization: slashing the cost of virtualization**  
*Technical Report PA005520, National ICT Australia, October, 2005*
- Ben Leslie, Peter Chubb, Nicholas Fitzroy-Dale, Stefan Gtz, Charles Gray, Luke Macpherson, Daniel Potts, Yueting Shen, Kevin Elphinstone and Gernot Heiser  
**User-level device drivers: achieved performance**

*J. Comput. Sci. & Technol.*, 20(5), 654–664, (September, 2005)

- Charles Gray, Matthew Chapman, Peter Chubb, David Mosberger-Tang and Gernot Heiser  
**Itanium — a system implementor's tale**  
*Proceedings of the 2005 USENIX Technical Conference, Anaheim, CA, USA, April, 2005*
- Peter Chubb and Darren Williams  
**Linux scalability — from the micro to the HUGE**  
*Proceedings of the 6th Linux.Conf.Au, Canberra, ACT, April, 2005*

# Our Hypervisors

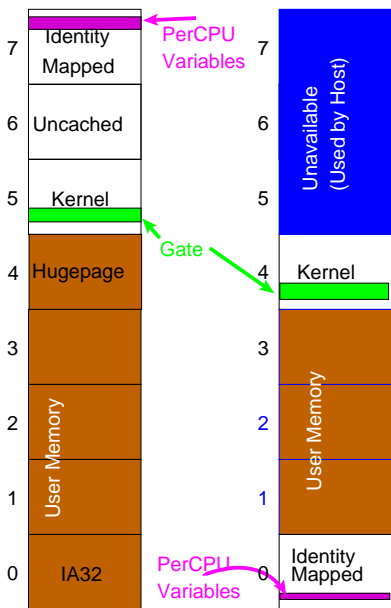
## Linux on Linux



- SKI simulator infrastructure used for device drivers.
- (future) allow drivers in guest to access real devices, using UserDriver framework.

- Uses **previrtualisation** and an address space rearrangement patch.

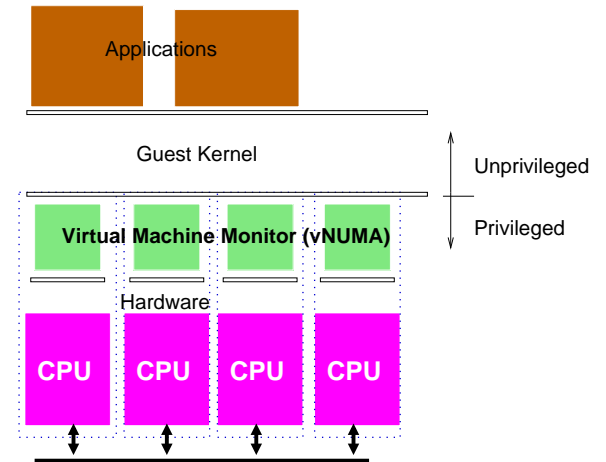
- Easy to extend to other guests, e.g., FreeBSD



- Host Linux uses regions 5–7.
- Guest kernel mapped into region 4. This precludes hugeTLBfs.
- Region 0 is the identity mapped region, for per-cpu variables. and where the hypervisor is mapped.

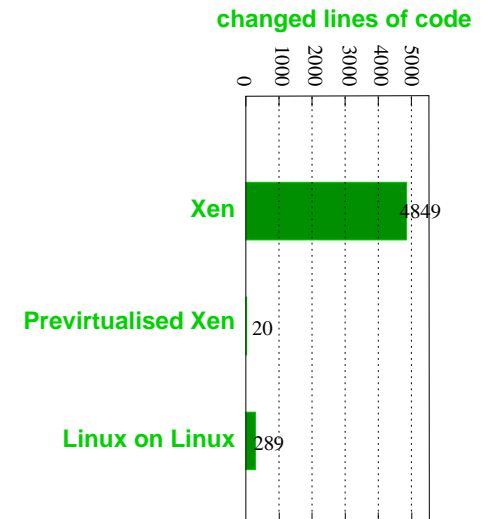
## vNUMA

- Make cluster look like ccNUMA.
- vNUMA VMM does distributed-shared-memory
- Commodity Gigabit Ethernet interconnect
- Guest Linux sees NUMA IA64 machine.



## Obligatory Xen section

- Xen requires **large changes** to guest kernel
- Use **previrtualisation** instead!
- Almost **no performance difference**.





# Gelato UNSW

Performance and  
Scalability on Itanium

[www.gelato.unsw.edu.au](http://www.gelato.unsw.edu.au)



THE UNIVERSITY OF  
NEW SOUTH WALES

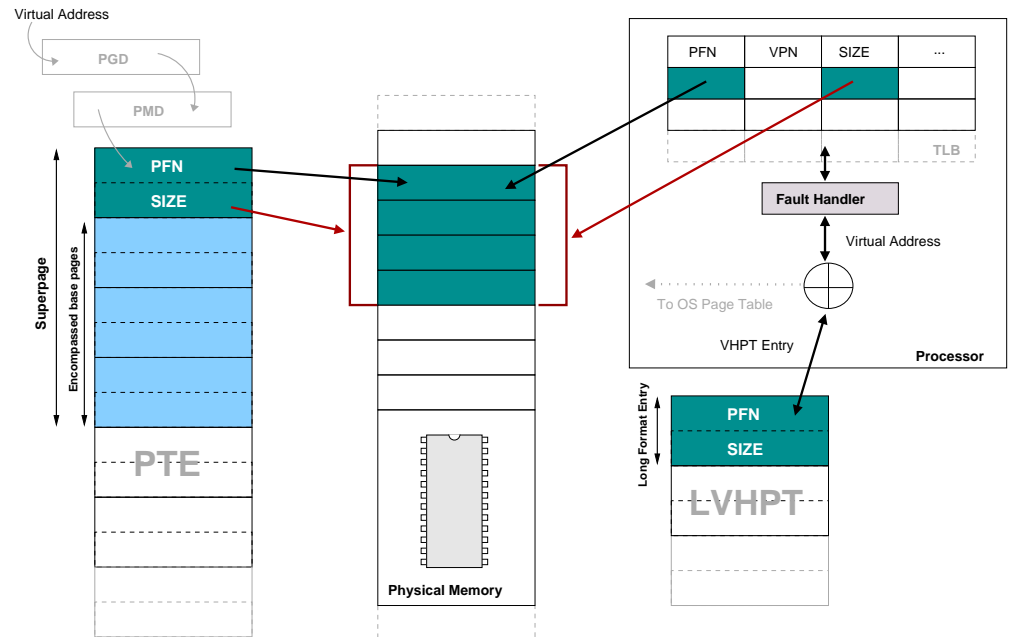


Australian Government  
Department of Communications,  
Information Technology and the Arts  
Australian Research Council

# Extending IA64 Linux for Superpages

## Infrastructure

- Virtual addresses on IA64 are translated via three mechanisms — all need support for multiple page sizes.
  - TLB.** The Itanium TLB has explicit support for page size in TLB entries. We have modified Linux to set page sizes when inserting into the TLB.
  - Virtual Hash Page Table (VHPT).** We have extended Linux to use the **long format** VHPT, which allows specifying a page size.
  - Page Tables.** We have doubled the size of a page table entry (**PTE**) to allow extra room for storing page size information.



## Research Questions

- Literature review of existing work.
- Superpage allocation policy (e.g. reservation based systems).
- Page migration, overall NUMA considerations.

# Virtualisation

To virtualise fully you need:

- **Clean separation** of system and user state.
- All Instructions that modify or inspect system state are **privileged**.

Real machines fall short on both these counts.

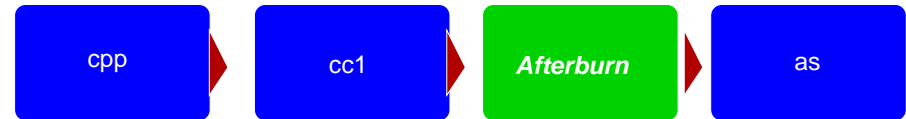
The choices are:

- Binary rewriting
- Previrtualisation — Modifying the guest via the toolchain
- Paravirtualisation — Manually modifying the guest operating system

## Itanium Virtualisation

- Itanium is generally easier to virtualise than x86,  
**But:**
  - Instructions like `cover` **non-privileged**, but behave differently in System mode
  - Instructions like `tash` and `ttag` reflect **non-virtual** system state but are non-privileged
- And trapping on illegal instructions is **slow**

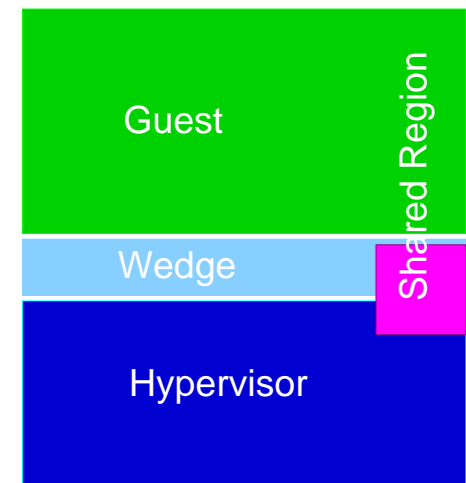
## Previrtualisation



In **previrtualisation** for Itanium we run a perl script (the **afterburner**) before the assembler. It transforms sensitive instructions into macros that call into the hypervisor via the **wedge**.

## Putting it together

The **wedge** either queries or updates the virtual CPU (in the shared memory region), or calls into the hypervisor. The wedge converts between the previrtualised guest and the interface presented by the hypervisor.



# Gelato@UNSW

## Who We Are

- Virtualisation (vNUMA),
- Reverse Engineering and performance tuning.

Matthew Chapman

### Paul Davies

- Page Table Abstraction,
- Guarded Page Tables,
- General Dogs-body.

### Peter Chubb

- Day-to-day team management,
- Virtualisation,
- User-level device drivers,
- Scalability.



### Ian Wienand

- Superpages

### Gernot Heiser

- Gernot leads the Gelato@UNSW team.

### Myrto Zehnder (Masters Student)

- From ETH Zürich
- User Level Drivers in a Virtual Machine

Gelato@UNSW is funded jointly by the University of New South Wales, National ICT Australia (NICTA), the Australian Research Council and Hewlett-Packard Company. SGI also support us

# Possible Future Work

## Scheduling

- Improved schedulers for high-end systems
- Hierarchical scheduling
- Per-process scheduling classes
  - Scheduling classes for ccNUMA
    - \* Topology scheduling
    - \* Gang scheduling within processor sets
  - Scheduling classes for multimedia
    - \* Isochronous
    - \* other soft-realtime schedulers
  - Scheduling classes for the Datacentre
    - \* Lottery schedulers
    - \* Fair-share and entitlement schedulers

## Virtualisation

- Linux on L4 (Wombat) on Itanium
- Exploring Vanderpool

## Network Performance investigations

- TCP/IP — Why so slow?
- More driver work

## More componentisation

### User level

- File systems
- TCP/IP and other protocol stacks
- More device drivers

### In kernel

- Processor schedulers
- Memory schedulers

This is part of our three year plan... not all of this will be done