

IBM Mainframe Hardware from a Linux Hacker's Viewpoint

General Architecture (1964-1980)

1964 S/360

- Defined by "Principles of Operation" document
- 16 x 32-bit registers
- 24-bit memory addressing
- SIO instruction starts channel I/O
- 2K or 4K pages in 1MB segments
- Per-physical page storage key and referenced/dirty bits

1970 S/370

- SIOF (Fast Release)
- DAT (Dynamic Address Translation)
- Dual address space instructions

1980 370/XA (eXtended Architecture)

- 31-bit addressing (2GB per address space)
- SIE instruction (Start Interpretive Execution)
- Dynamic channel reconnection

General Architecture (198x-2001)

198x 370/ESA (Enterprise Systems Architecture)

- SSCH instruction starts subchannel I/O
- 16 address space access registers
- Extra page moving instructions
- PR/SM (LPAR - logical partitions)

1990 ESA/390

- IEEE-compliant Floating Point
- Suppression-on-protection (for better page fault handling)
- Extra immediate and relative instructions
- ESCON

2000 (z/Architecture)

- 64-bit superset of ESA/390
- SIE in 31-bit or 64-bit
- Principles of Operation released Jan 2001

Processors and their uses

Each PU (IBMese for CPU) can be used as a

- CP (Central Processor) which runs ordinary code
- SAP (Service Assist Processor) runs I/O microcode
- ICF (Internal Coupling Facility) runs clustering code
- Spare (available for expansion or sparing)
- n-way (in model numbers and elsewhere) only counts CPs

Every PU has dual execution units and a comparator

- A PU failure causes a retry
- A retry failure causes a transparent transfer to a spare PU
 - IBMese for this is sparing

Current mainframe models

G5 and G6 (G for Generation) are modernish

- 9672-X46 is a 4-way G5 (sic)
- X97, XX7, XY7, XZ7 are 9, 10, 11, 12 way G6
- Replace initial X with Z for "Turbo" (faster clock)
- 14 PUs (e.g. 12 CPs + 1 SAP + hot-spare)

z900 is the new 64-bit model just released

- z900 (2064) has models 101-116, 1C1-1C9 (1-way to 16-way)
- 101-109: maximum 12 PUs (9 CPs + 2 SAPs + hot-spare)
- 110-116: maximum 20 PUs (16 CPs + 3 SAPs + hot-spare)
- 1C1-1C9: 1-9 CPs but can expand all the way to 16

Multiprise 3000 (models H30, H50, H70)

- Low-end, cut-down models without all RAS features of real mainframes
- 4 PUs, CP+SAP: 1+1 (H30/H50), 2+1 (H70)
- Internal DASD (via RAID5 SSA 18GB disks)
- Max: 56 channels, 792 GB DASD (visible)

MCM: the heart of the system

The heart of a system is the MCM (Multi-Chip Module)

- A single large module (~ 12cm x 12cm) which includes
- All PUs (all present, disabled/enabled by licensing code)
- L1 cache for each PU
- Binodal L2 cache (means half PUs have one, half have another)
- SC (System Controller) chip
- MBA (Memory Bus Adapter) chips
- SC, MBA, L2 and all buses clocked at half microprocessor speed

L1/L2 cache

L1/L2 caches are

- big in total size, associativity and cacheline size
- reliable (parity in L1, ECC and sparing in L2)
- attached to CPUs and memory with very high bandwidth (16GB/s for G6, 24GB/s for z900)

Hardware-assisted move engine

- uses line buffers in L2 chips
- knows which memory banks are active

L1/L2 cache comparison

System	Cache line size	L1 size (I+D)	L1 assoc	L2 size	L2 assoc/shared	Clocks in MHz for		
						CPU	memory	I/O
PentIII/Xeon	16	16K+16K	2?	256K-2MB	2?/no	800	133	33/66
S/390 G6	256	256K	4	8MB	8/yes	637	318	318
z900	256	256K+256K	4	16MB/32MB	8/yes	770	385	385

Channel subsystem (CSS)

Channels

- A channel is a processor connected to main memory and one or more control units.
- Current systems support 256 channels.
- Identified by an 8-bit channel path id (CHPID)

Control units

- A control unit (CU) is attached to up to 256 devices, e.g. DASD (IBMese for disks) or tapes.
- Sometimes a separate box; sometimes shares physical box with device

Each device is identified to

- humans by a 16-bit device number
- the CSS by a 16-bit subchannel number

Channel subsystem (contd)

Subchannel number

- indexes into an array of UCWs stored in the HSA (Hardware Storage Area), memory available only to microcode.

UCW (Unit Control Word) includes

- device number
- Unit Address (UA), 8-bit identifier within channel
- channel path information

Channel path

- Each device can attach to up to 8 different channels for reliability and performance
- Devices may connect to more than one CU

Channel programs

Channels execute channel programs

- Channel programs consist of CCWs
- CCWs are Channel Command Words
- A mini machine-language for I/O which can
 - read/write between memory and devices
 - do device-specific control functions
 - test device/channel status and loop

Channel subsystem instructions

Single instruction (SSCH, Start SubCHannel) initiates I/O

- operands are
 - subchannel number
 - ORB (Operation Request Block) address
- ORB includes channel program address

SAP takes over

- copies ORB into UCW
- places UCW in initiative queue in HSA
- finds appropriate channel path
- gets channel to execute channel program

When I/O is done

- status is stored in UCW
- SAP generates an I/O interrupt

QDIO (I/O for OSA-Express network) bypasses CSS

STI (Self Timed Interconnect) bus

STI bus

- connects to main memory
- Physically appears as I/O slots
- Maximum cable length: 10m
- other buses connect to STI bus

G6 STI

- Bandwidth: 333 MB/s (bidirectional)
- Limit: 16 (?), all concurrent

z900 STI

- Bandwidth: 1 GB/s (bidirectional)
- 24 of them, 6 on each MBA (all concurrent)
- Up to 36 physical links

ICB (Integrated Cluster Bus)

- attaches to STI for cluster (Parallel Sysplex)

FICON channels

FICON

- Fibre-channel (FC-PH Standard at physical layer)
- Bandwidth: 100 MB/s
- I/Os per second: 3600 on G6, 4800 on z900
- Limits: 36 on G6, 96 on z900
- Maximum cable length: LX: 10km (20km via RPQ, 100km with repeaters); SX: 500m

Connects to

- Not much native yet (only a 3590-A60 tape controller)
- 9032 Model 005 ESCON director (16 ports)
- New FICON-only directors (32-port and 64-port)

ESCON and parallel channels

ESCON

- The main way of connecting devices (apart from older systems)
- Fibre is the physical medium
- Bandwidth: 20 MB/s (17MB/s for data)
- Limit: 256 (subject to maximum of 256 channels)
- Maximum cable length: 9km

Parallel channels

- Historical now (inevitably still in legacy use)
- "Bus and tag" (two large cables per connection)
- Bandwidth 4 MB/s, maximum length 400 feet

OSA Network

OSA-Express

- Has different versions
 - GbE: Gigabit ethernet at 1 Gbps
 - FENET: Fast ethernet at 100 Mbps
 - 155 ATM: ATM at 155 Mbps
- z900 GbE version
 - 64bit/66MHz PCI
 - can reach line speed
 - has hardware support for lots of stuff
- Uses QDIO (not CCWs) for I/O

OSA-2

- For G6 and earlier or z900 in compatibility cage
- Versions for 10 Mbps ethernet, FDDI, token ring, ATM, SNA and other weird stuff

Parallel Sysplex (clustering)

Parallel Sysplex is IBM's name for mainframe clustering

At least one PU acts as a CF (Coupling Facility)

- Can be an ordinary PU
- Can be a PU in a 0-way box (e.g. z900 model 100)

ICB (Integrated Cluster Bus)

- attaches to STI for cluster
- G6 ICB: 333 MB/s; z900 ICB-3: 1 GB/s
- Maximum cable length: 7m
- Limits: G6: 18 ICB; z900: 16 ICB-3

ISC (InterSystem Coupling)

- 4-port ISC card attaches to STI bus
- G6 ISC: 1 Gb/s; z900 ISC-3: 2 Gb/s
- Max cable length: 10km (20km via RPQ at 1 Gb/s)

PCICC (PCI Crypto Coprocessors)

Up to 16 crypto coprocessors can be added

- 8 PCICC cards, each containing two processors
- G6 has two CMOS crypto coprocessors built in

Algorithms

- DES
- 3DES
- RSA
- SET and others

Secure

- Tamper proof
 - Zeros memory if too hot/cold/tampered
- Hardware key entry

z900 can do 2000 SSL transactions per second

Physical form

One main frame

- Called the A frame
- Half is the CPC (Central Processor Complex)
- Half is a cage of I/O slots (22 in G6, 28 in z900)

One expansion frame

- Called the Z frame
- Two cages of I/O slots
- For z900 choose between
 - nIO (new I/O cage) for new z900 cards
 - cIO (compatibility I/O cage) for older cards

Optional B frame (small width)

- For IBF (Internal Battery Facility), built-in UPS
- Main supply is 2 x 3-phase: 1 x 2-phase in normal use, rest for redundancy

RAS: Reliability, Availability and Serviceability

ECC throughout (L2, memory, all data buses and channels)

- L1 is parity, write-through, write-back-managed (think about it :-)

Retry, correct, recover, sparing, checkstop at finest possible granularity

G5 SC chip has >1500 fault detectors; L2 chip has >250

Processor sparing--transparent wherever possible

Memory error detection/correction/sparing

I/O errors

- Retried, corrected where possible
- Redirected via a working channel if necessary
- Hardware that goes mad and spews interrupts is "boxed"
- All of the above is transparent

All errors are logged and traceable in gory detail

Upgrading

PlanAhead

- Capacity Upgrade on Demand (CUoD): concurrent (i.e. hot)
- Concurrent Conditioning: add stuff in advance; pay less

Concurrent upgrades for

- CPUs can be enabled quickly up to MCM limit
- I/O (and crypto) cards can be hot-added on z900 only
 - fast turn-on of existing on-card channels on any system
- Memory
 - up to existing on-card memory (e.g. multiple of 4GB)
 - intention to introduce hot-add of memory on z900

Capacity Backup Upgrade (CBU)

- Put extra stuff in a system intended for disaster takeover
- Pay less than usual and normally run without it
- When disaster hits, phoning IBM gets key to enable extra capacity
- Disaster recovery provision allows ~2 weeks/year free testing

Splitting up the hardware and software

A hypervisor is a special purpose O/S which creates a virtual machine environment for its guests

VM (Virtual Machine) is IBM's canonical example

- current versions VM/ESA for S/390 and z/VM for z900
- (expensive) software but extremely flexible

LPAR (Logical PARTitions)

- cut-down hypervisor comes with every S/390
- maximum 15 partitions; some changes not concurrent

VIF (Virtual Integrated Facility)

- introduced specially for Linux/390
- very basic functionality compared to VM
- ...but much cheaper
- ...and without the 15 guest limit of LPARs

VM

A text file describes a virtual machine configuration for each userid

When a userid logs in, the virtual machine is created

- with its default amount of memory
- a virtual console, printer, card reader, card punch
- disks backed by real whole disks or chunks of real disks
- tapes, channel-to-channel links (for TCP/IP)

the user types commands to CP (Control Program)

- can query and reconfigure the virtual machine
- can query and make links/changes to other virtual machines (if userid has appropriate privileges)
- can IPL (Initial Program Load, i.e. boot) a guest O/S such as Linux/390, CMS, OS/390 (or even VM)
- CP is friendly and initial login commands can be automated

VM/CMS

CMS is a single-user O/S

- but run under VM, each user has a separate virtual machine
- and CMS has some support for sharing files and devices

The whole system can be administered with CMS and CP

- logged in as a special user
- MAINT user analogous to root but privileges can be split further
- a text file configures users/virtual machines
- using CMS and CP to manipulate and store information from
 - the real hardware and virtual machines
 - real and virtual devices
 - a large amount of auditing and performance information

VM Performance

Rules of thumb

- 6 MB RAM per idle Linux guest image
- 0.3% of one CPU per idle guest ("300-350 per engine")
- 1 in 12 guests active at any time (hand-waving)

David Boyes "Linux Scalability: Test plan Able/Baker/Charlie"

- Test plan Able/Baker: 250, 2750, 10000 images
- Test plan Charlie: add images until bad things happen
 - Linux tweaked (lower HZ) to coexist better
 - each image ran Apache, INN, DNS or client tester
 - at 30000 images, still got subsecond response
 - at 41400 images, CP refused gracefully to add another image

A reasonable number of Linux guests is "hundreds or thousands"

VM has very good performance monitoring and tuning functionality

VM Examples I

Administrator gives Fred a new disk (say devnum 2000)

- Admin does: ATTACH 2000 TO FRED AS 2000
- or just ATT 2000 FRED *
- To use new disk under Linux, Fred can
 - reboot his Linux guest
 - or use dynamic device support in recent Linux
- don't forget to keep user directory in sync

Fred and Bill want a private network connection between themselves

- Fred does: DEFINE CTCA AS 1000
- Creates a virtual Channel to Channel Adapter
- Append USER BILL to restrict coupling to Bill
- Bill does: DEFINE CTCA AS 1001 to create his own
- then: COUPLE 1001 TO FRED 1000 to connect them

VM Examples II

Fred wants a memory upgrade

- shutdown -h +5 "Linux can't add memory concurrently, yet"
- DEFINE STORAGE 128M
- IPL 131 (or wherever his Linux kernel lives)
- Now Fred has 128 MB of RAM
- User directory holds default amount and maximum allowed

Fred wants a temporary 1 GB disk

- DEFINE T3390 AS 192 CYLS 1500
- Now Fred has 1500 cylinders of disk on virtual device 192

Fred wants a few more CPUs

- DEFINE CPU 1 2 3 adds CPUs 1, 2, 3 to go with CPU 0
- They are only virtual but
 - may help with scheduling or debugging
 - or be used to distribute resources better
- Administrator can dedicate real CPUs or memory

VM Examples III

Administrator distributes CPU and paging resource shares

- SET SHARE FRED ABSOLUTE 20%
 - Fred is guaranteed 20% whatever happens
- SET SHARE BILL RELATIVE 100
- SET SHARE BOB RELATIVE 100
- SET SHARE ALICE RELATIVE 300
 - Bill and Bob compete with same relative share
 - Alice gets three times as much
- Also have
 - hard and soft limits
 - special tweaks for interactive scheduling
 - lots of global settings available to monitor and twiddle

Administrator can throttle I/O per device (though not per user)

Limits

Channels: 256

CPUs

- G6: 14 (12 CPUs + 1 SAP + 1 spare)
- z900 101-109: 12 (9 CPUs + 2 SAPs + 1 spare)
- z900 1C1-1C9,110-116: 20 (16 CPUs + 3 SAPs + 1 spare)

Memory

- G6 and z900 101-109: 32 GB
- z900 1C1-1C9,110-116: 64 GB

FICON

- G6: 36
- z900: 96

OSA (ethernet, fast ethernet, gigabit ethernet etc.)

- G6: 12
- z900: 24

Various interrelating device-related limits (mostly large)

- 65536 device numbers; 80000 subchannels (G6)
- 8 paths per device; 256 CHPIDs; 256 devices per channel

Could do better

Prices aren't public

VM I/O resource control could be better

Disk subsystems are restricted

- Not many vendors do channel attached disk
- 390 OSeS mostly rely on CKD instead of FBA
- CKD (Count Key Data): disk consists of (Key,Data) records
- FBA (Fixed Block Addressing): disk has fixed sized sectors

Prices are high

FICON is not widely supported natively yet

FICON does not interoperate with other FC (above frame layer)

ESCON is only 17 MB/s (though you can have many of them)

Non-390 I/O adapters are much higher performance (1 for 1)

Did I mention prices?

Much of the available online information

- contains biased marketing not technical comparison
- still ignores Linux
- confuses (or lumps together) OS/390 with S/390