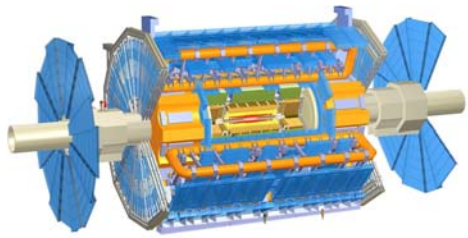
Plumbing at 10Gbps

The art and science
of high speed data transmission

Brian Martinescu
CERN
Geneva

The ATLAS Experiment



The collision volume is surrounded by detectors and a magnetic field to bend tracks to enable momentum calculations

Data Moving Problem

- Detector produces analogue signals
- Conversion to digital close to detector
- Processing is done ‘far’ away
- Data starts life in a silicon chip
- Moved to a cable
- Received in a chip
- Electrical connectivity assured
- So what’s the problem?

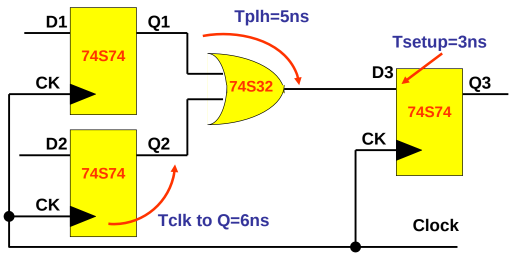
History   Lesson

Once upon a time, many many years ago
Digital Electronics were for MEN

YES	NO
RIGHT	WRONG
ON	OFF
ONE	ZERO
TRUE	FALSE

It was the golden age of
Logic, Truth Tables, State Machines and Rock and Roll

Building Logic was a really easy job
Wire logic blocks together
Add up the delay times from the data sheet
And work out the Clock Timing budget.



Clock Timing budget :

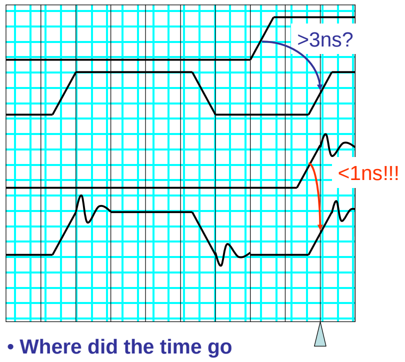
6ns +5ns+3ns=14ns. Clock max= 71.43MHz

Engineers are conservative so run circuit at 70Mhz

We power up and run the tests and --------

-------- it doesn't work!

Get out the oscilloscope!



- **Where did the time go**
- **What's all that ringing?**

Where did the time go?

It's all Einstein's fault!

Speed of light C=30cms/ns in a vacuum

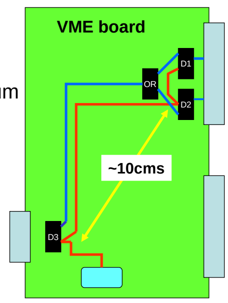
An electric field through fiber glass :

15 to 20 cm per ns

What's all the nasty ringing noise?

It's all Maxwell's fault!

Propagating signals reflect from impedance mismatches.



And all this is at less than 100Mhz!

My logical world is coming to an end

It's time to go back to school!

Maxwell's Equations

Integral form in the absence of magnetic or polarizable media:

I. Gauss' law for electricity $\oint \vec{E}\cdot d\vec{A} = \frac{q}{\epsilon_0}$

II. Gauss' law for magnetism $\oint \vec{B}\cdot d\vec{A} = 0$

III. Faraday's law of induction $\oint \vec{E}\cdot d\vec{s} = -\frac{d\Phi_B}{dt}$

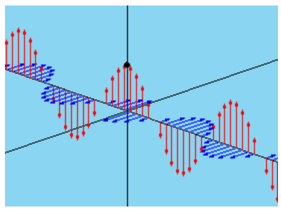
IV. Ampere's law $\oint \vec{B}\cdot d\vec{s} = \mu_0 i + \frac{1}{c^2}\frac{\partial}{\partial t}\int E\cdot dA$

Wave Propagation

Definitive description is in Maxwells equations

Here they are just so you can say you've seen them

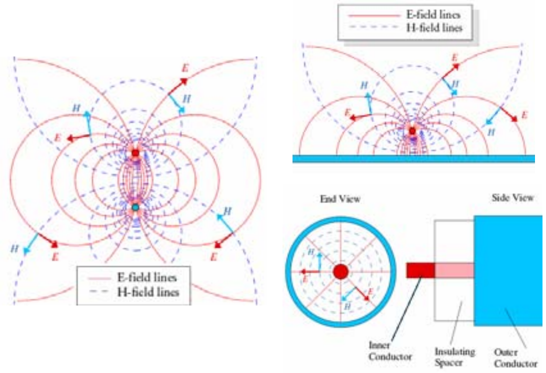
Solving them depends on the boundary conditions. Kelvin wanted to see a mechanical model, and so do I



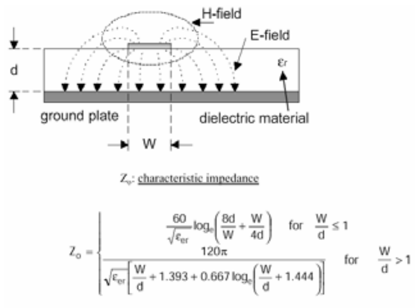
Propagation in free space

E H

E H solution for conductors

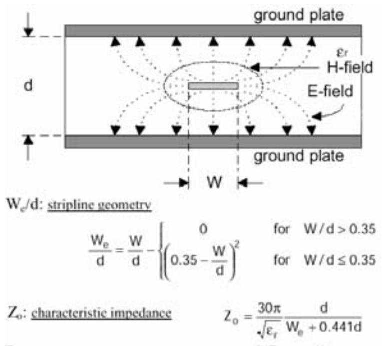


Microstrip structure

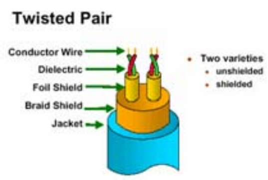


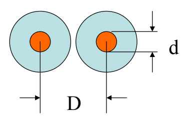
Stripline Structure

Easy to build in PCB's



Twisted pair

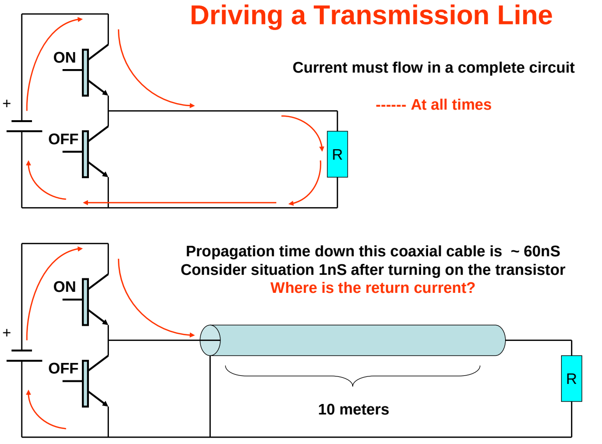




$$Z = \frac{120}{\sqrt{\varepsilon_r}} \ln\left(\frac{2D}{d}\right)$$

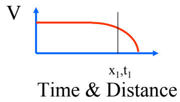
Where's the the twist variable??
Beware of simplifications

Driving a Transmission Line

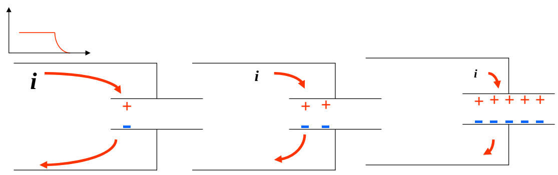


Wave Propagation

Applying a step voltage to a conductor generates an electro-magnetic wave that propagates down that conductor.
This plot shows the state at some time t_1
The wave front is building up at x_1.
Electric field is increasing
Magnetic field is increasing
This takes time.
Building up an electric field between a conductor and an adjacent conductor or plane: this looks like capacitance
Building up current creating a magnetic field: looks like inductance



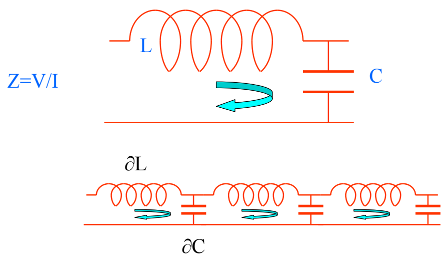
What is a capacitor?



Current flows
charge is separated
voltage rises
current diminishes

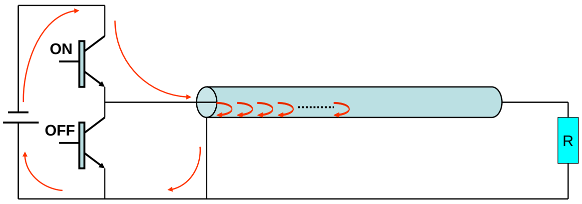
As the capacitor charges, return current flows

A Transmission Line Model

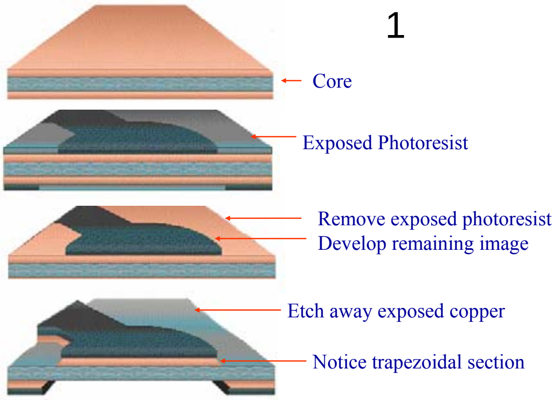


As the wave front advances down the transmission line there is a continual return current generation from each incremental L/C slice

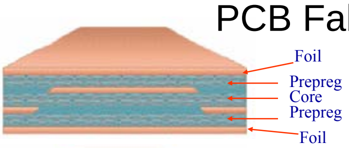
Wave Propagation



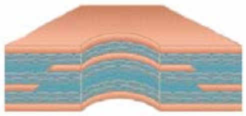
PCB Fabrication: 1



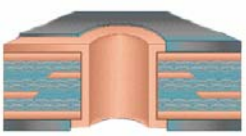
PCB Fabrication: 2



Lamination stage:
- assemble,
- press together
- cure at high temperature

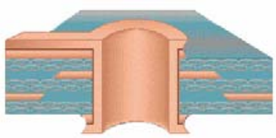


Drill the through holes then deposit a thin film of copper on all surfaces.
The deeper the hole the wider it must be for uniform plating

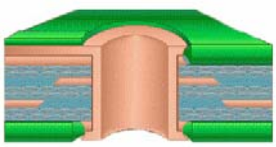


Add a layer of resist and develop it so that there is no resist where you want copper to be.
Then deposit copper on tracks and vias

PCB Fabrication: 3

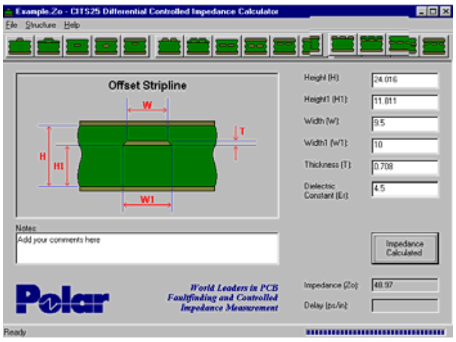


One more etch step to remove the remaining copper exposing the surface tracks

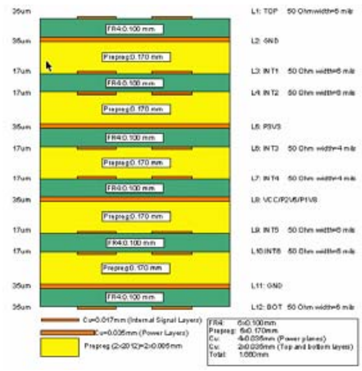


Add solder resist film to protect areas that are not to be soldered.
This affects the track impedance

Building a transmission Line Calculating Impedance



Typical PCB Stack



The prepreg layers are made up of many layers of different thicknesses

All of different dielectric constant!

Note how narrow the the 50ohm tracks are.

Raw Material: Tolerances

Laminate Thickness & Tolerances

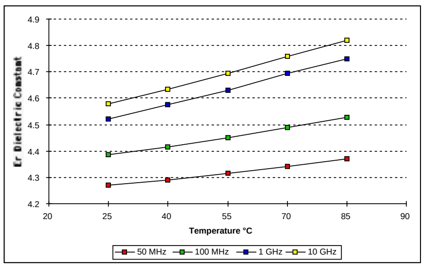
Core Thickness Range(inches)	Tolerance Class 1	Class 2	Class 3	Class 5
.0016 to .0045	±.0010	±.0007	±.0005	-.0005 + .0010
.0046 to .0065	±.0015	±.0010	±.0007	-.0007 + .0012
.0066 to .0120	±.0020	±.0015	±.0010	-.0010 + .0015
.0121 to .0199	±.0025	±.0020	±.0015	-.0015 + .0020
.0200 to .0309	±.0030	±.0025	±.0020	-.0020 + .0025
.0310 to .0409	±.0065	±.0040	±.0030	-.0030 + .0035

Prepreg Resin Content

Thickness	% Resin
1.7mils	75±3
2.5mils	66±3
3.5mils	56±3
4.4mils	56±3
5.0mils	50±3
6.9mils	43±3

Copper Cladding
18 micron ±10%
35 micron ±10%

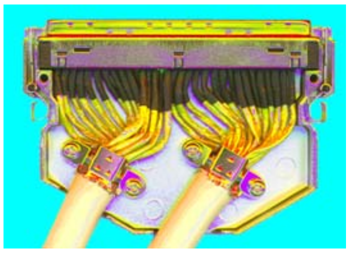
FR4 Dielectric Constant vs temp

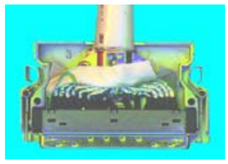


10% impedance match economically achievable

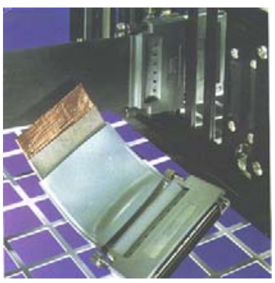
Coaxial connectors

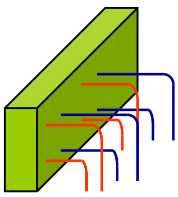
Connectors: 68100 Pin 3M



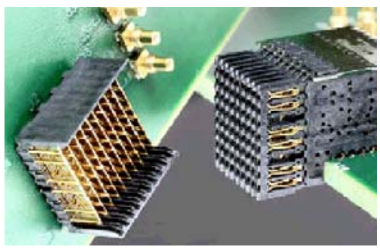


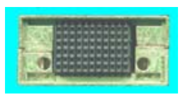
Pleated foil cable





Connectors: VHDM





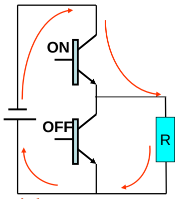
Plumbing the chip

Q: What happens when a link driving output transistor turns ON?
A: Current flows.

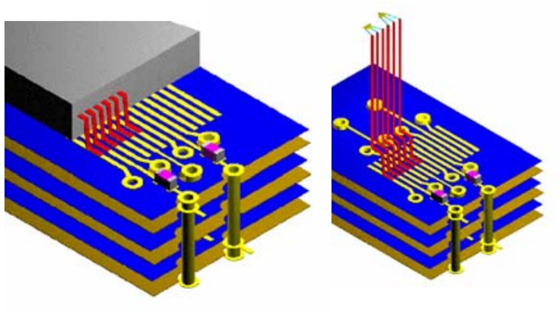
Q: Very good, but where does that current come from?
A: The power supply

Q: Sure, but it takes time for the supply to know it has to deliver a bit more current. What happens in the meantime.?
A: Err, umm, The voltage drops?

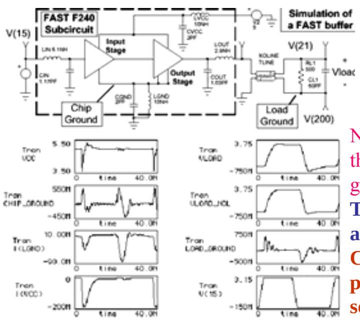
Unfortunately, yes it does, and then worse things start to happen!



Chip to PCB



Ground bounce model



Note inductance on power, ground and signal lines.

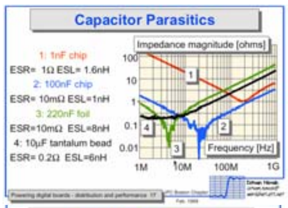
Note nasty effects on the chip's local power and ground.

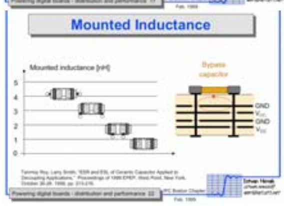
The output signal suffers a little but the worst is:
Consider the effect of this power supply driving your sensitive receiver on the same chip!!!

Capacitors = capacitance?

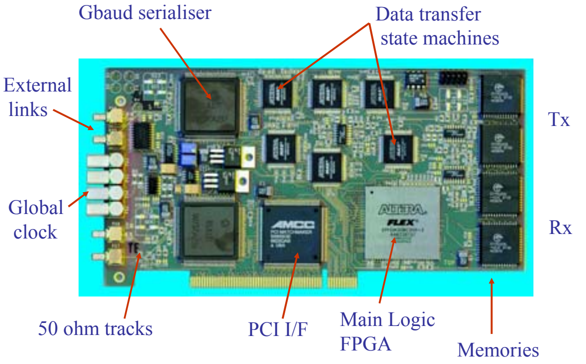
Capacitors behave like inductors above a certain frequency.
Need to supply more than one per power pin to cover the range of supply fluctuations: Core, I/O, Serial I/O

Mounting capacitors on a PCB
Increases the inductive component inherent in the device.



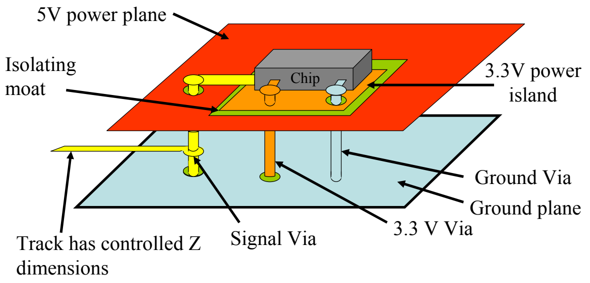


NIC

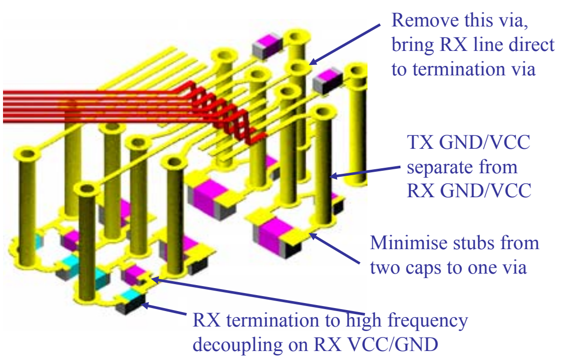


What is wrong with this picture?

When there are several supply voltages it is sometimes necessary to embed one power plane within another to reduce number of layers.

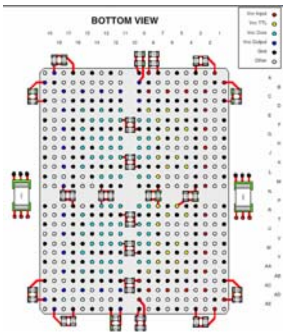


Flat Pack Decoupling



De-coupling a AMCC 34*34 @ 3.2Gbps Cross-point switch

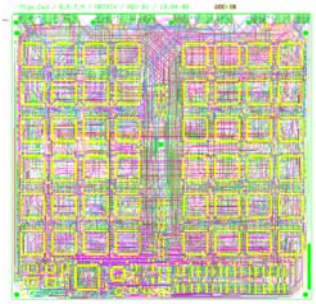
On a BGA package there is less inductance in the package but much less room to place de-coupling for more than a few of the power pins



Typical complex design

Once the design rules are understood.

Complex networks can be achieved



Introduction to TDR

Mismatched printed circuit traces, connectors, stubs and loops in decoupling circuits, all are sources of ‘extra’ L and C in the circuit.

How much is there, what effect do they have, and when do we care?
You can measure it by injecting a pulse and watch for the echos.

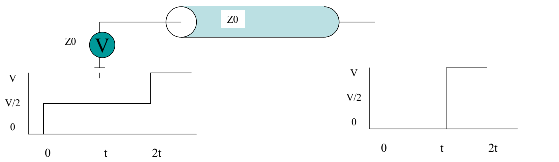
Q: Why are there echoes?

A: Impedance mismatch.

Return to the transmission line theory.

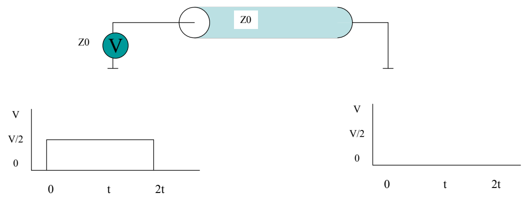
Extreme case of high Z mismatch

If a wave propagating between delta sections of Z meets an open:
- there is nowhere for the energy to go
- the incoming current builds up potential on the last delta C
- the potential rises
- it starts to propagate a rising wave in the opposite direction

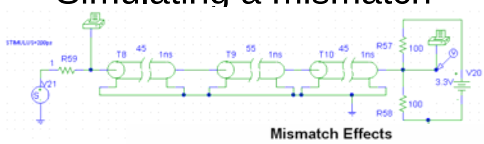


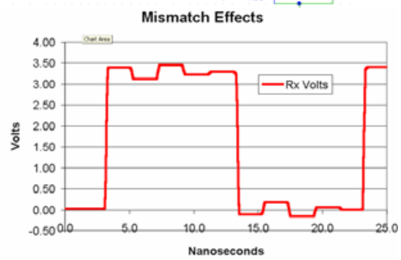
Extreme case of low Z mismatch

If a wave propagating between delta sections of Z meets an short:
- all the incoming current is shorted to ground
- the potential falls to zero
- it starts to propagate a falling wave in the opposite direction



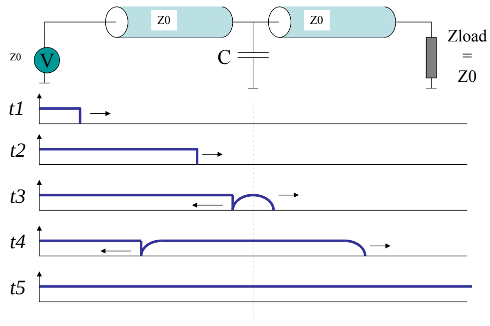
Simulating a mismatch



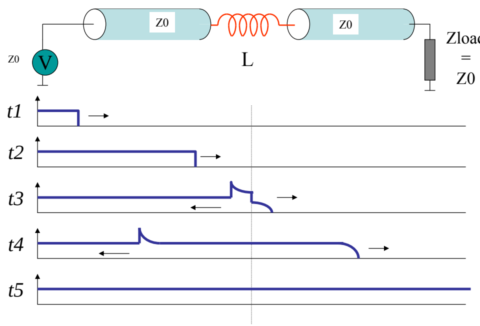


Mismatched lines
50Ω ±10%
Multiple reflections
Lifts logical 0
Reduces logical 1

Effect of excess C

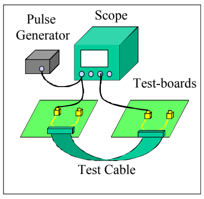


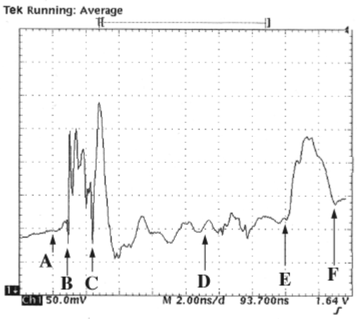
Effect of excess L



Go to animations

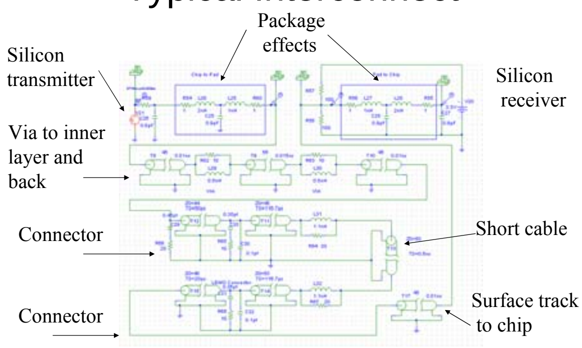
TDR technique





Inject a fast pulse into a transmission network
Measure the transmitted and reflected waves
Extract equivalent circuit
Use model for simulations.

Typical Interconnect



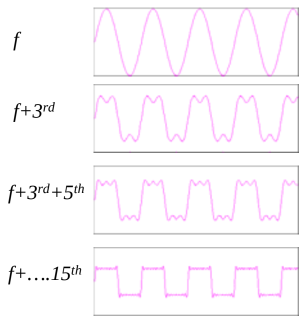
Simulated Results

200ps rise time pulse

Tr=291ps
DC Vdiff=1.6v
AC Vdiff=1.2v

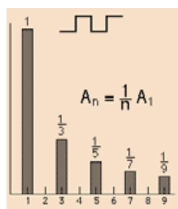
If we increase pulse edge to 100ps then Tr=219ps

Frequency and Fourier



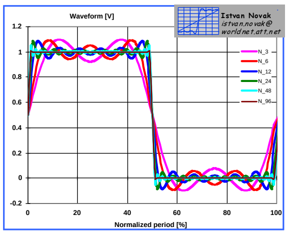
Even with the fifth harmonic we don't yet have a good approximation to a square wave.

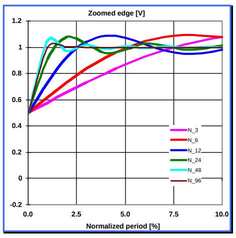
Fast edges introduce ***high frequency components***



Harmonics and rise time

It's not too much to ask is it:
10Mhz, 50% duty cycle with a 1ns rise time?
No problem: Provided you supply some GHz component!



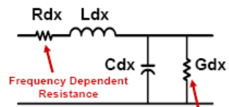


How fast can we go?

Board Losses

- Skin Effect Causes Frequency Dependent Series Resistance on the Line
- Dielectric Absorption Causes Conductance of the Line to Vary with Frequency
- Both the Skin Effect & Dielectric Absorption Will Increase Attenuation At Higher Frequencies

Lossy Model

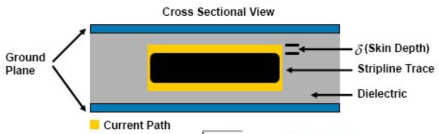


$$Zo = \left(\frac{R + Ls}{G + Cs} \right)^{1/2}$$

Frequency Dependant Conductance

Skin Effect

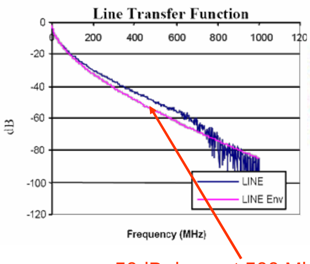
- High Frequency Current Flows Primarily on the Surface of A Conductor
- Changing Current Distribution Causes Resistance to Increase As Function of Frequency



$$\delta = \sqrt{\frac{\rho}{\pi F \mu}}$$

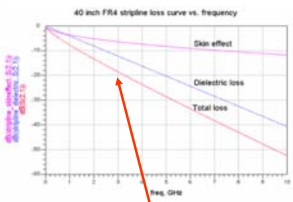
Resistivity
Permeability of Free Space
Frequency

100 meter cable



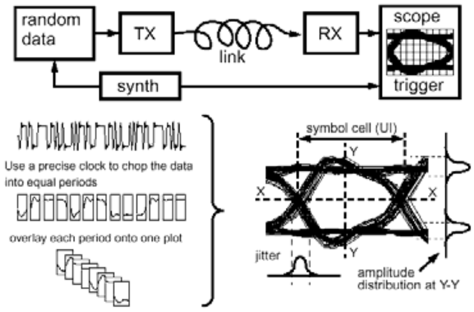
50dB down at 500 Mhz

40 inch backplane

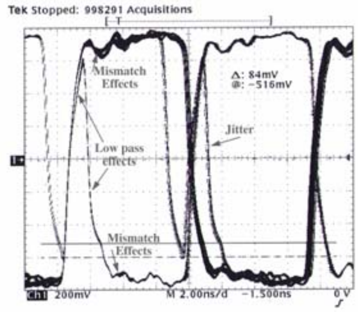


20dB down at 3 Ghz

Small Diversion: The Eye Diagram



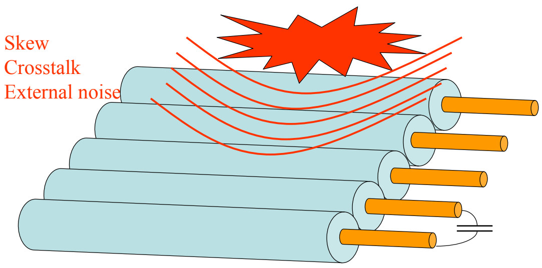
HSlink over PCB and coax



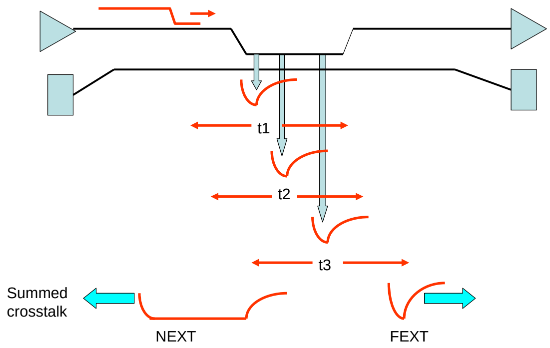
Conclusion
Definitely going to have to work hard to manage over a long length of cable
Might find a fairly simple solution for short haul over a backplane

Back to the 10G problem

- It's fairly obvious we have frequency problems
- Can we "divide and rule' by using multiple parallel paths

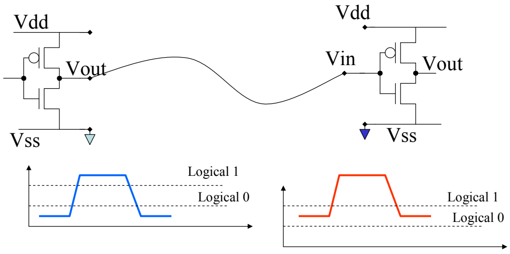


CROSSTALK



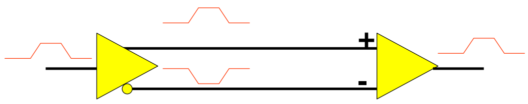
Driver to receiver problem

Single ended driver/receiver



Single ended transmission sensitive to noise and offset in remote ground levels

Differential Transmission



Transmitter sends the data signal and its negative
Receiver is now a differential amplifier
Very sensitive to small difference between the two inputs
Very insensitive to common mode noise
BUT now we need twice as many signals in the interconnect!

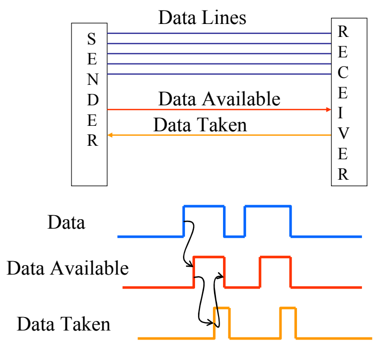
The story so far. We have seen that

- we need to have well engineered transmission lines
- there is a physical bandwidth limitation
- you need a lot of bandwidth to obtain ‘clean’ signals
- care must be taken with connectors and cables
- differential transmission is the best point to point solution

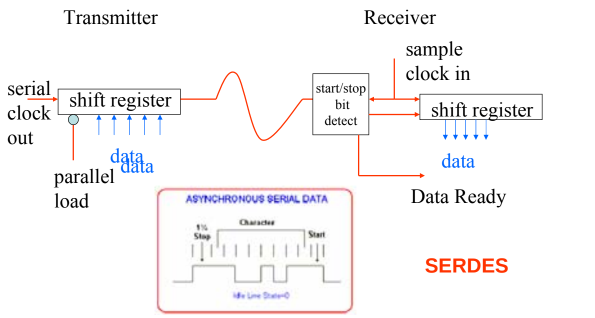
BUT

That was all just for a single channel of ‘bits’
What about timing information?
What about control information?
We look now at bit sequence transmission

Transmitting Data
Parallel Asynchronous



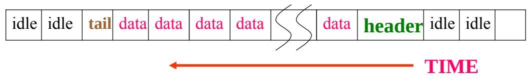
Serial Asynchronous



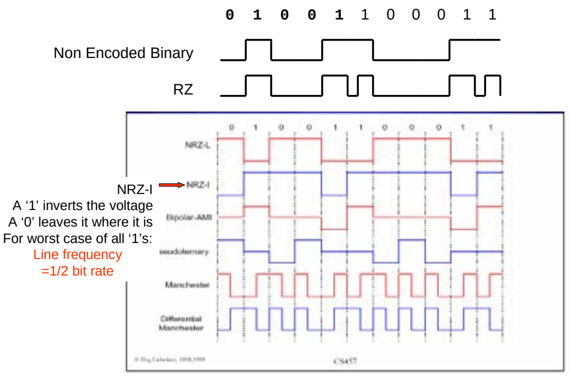
Serial Synchronous

With asynchronous serial
- can tolerate up to 5% difference in Tx/Rx clocks
- overhead of start and stop bits

Send the clock embedded in the data stream.
The receiver locks on to the clock and samples the data bits
Must send something even when there is no data
Choose a special pattern of bits, send it at the beginning and when there is nothing to send.
Start a frame with a **header** holding address or /command
Send the 'payload' of **data**
End the frame with error checking information in the **tail**



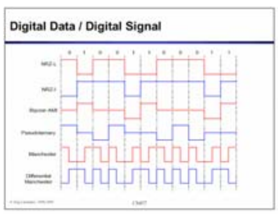
Line Encoding



Defining transition algorithm is not enough

Need to restrict bit sequence options:
Not too many consecutive zeroes
Not too many consecutive ones
Otherwise:
- Lose clock synchronisation
- DC content in data stream

Encode Data



8B10B

- 10 bits used to encode each octet
- Each code group must have:
  - 5 ones and 5 zeroes
  - Or 4 ones and 6 zeroes
  - Or 6 ones and 4 zeroes

Running disparity calculation
Maintains DC Balance
Maintains easy clock synchronisation
Adds overhead:
for a 1Gb/s 'payload we need to send 1.25Gb/s

Spectral response of codes

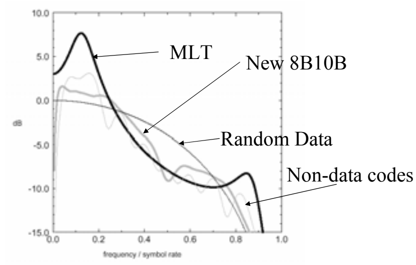


Figure: Power spectral density

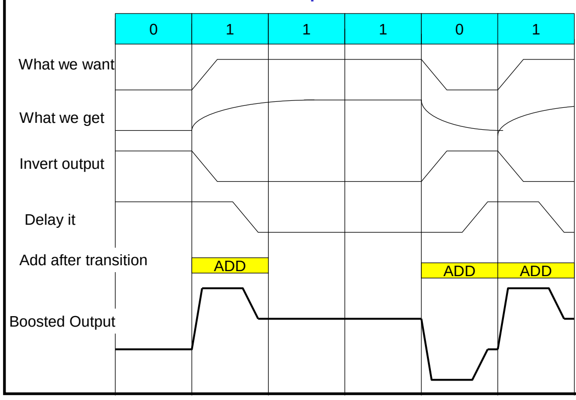
Sample of data codes

1111111111
0101010101
1111100000
1100110011
0001100111
1110010010
1111110000
0001110011
0011000111

10Gb/s for a Backplane?: do the math

- 10Gb/s payload +10/8 overhead=12.5Gb/s
- Divide by two: 6Gb/s no way **(Oh really?)**
- Divide by four: >3Gb/s ???? Well maybe
- 3.125Gb/s per channel
- NRZ gives a bandwidth need of ~1.6 Ghz
- Haven't got it for backplane distances
- High frequency losses
- What if we could boost the high frequency components
- More power, more cross talk, more interference
- But a backplane is a 'sheltered' environment + lots of power

Pre-emphasis



PRE-EMPHASIS EFFECTS ON SIGNAL INTEGRITY

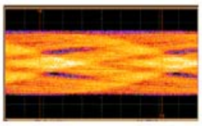


Figure 3: 3.125 Gbps Signal Sent Over a One-Meter FR4 Backplane Plus Two Connectors (Pre-Emphasis Turned OFF)

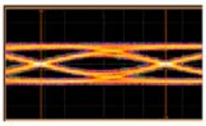


Figure 4: 3.125 Gbps Signal Sent Over a One-Meter FR4 Backplane Plus Two Connectors (Pre-Emphasis Turned ON)

Figure 5a 800mV signal with pre-emphasis, on operating at 3.125Gbps, taken at 0".

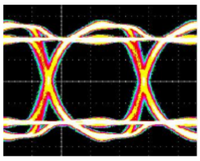
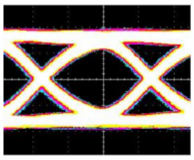


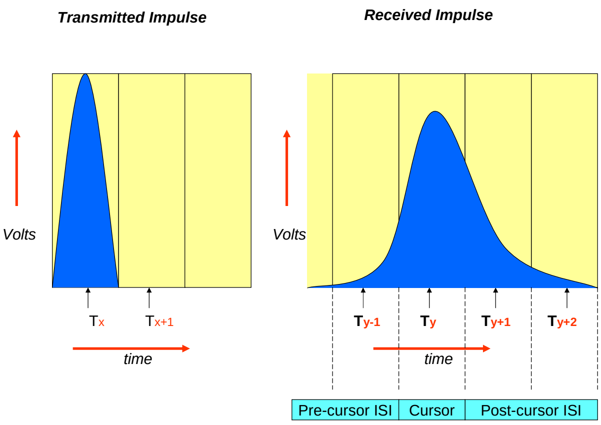
Figure 5b 800mV signal with pre-emphasis, on operating at 3.125Gbps. Taken at 20"



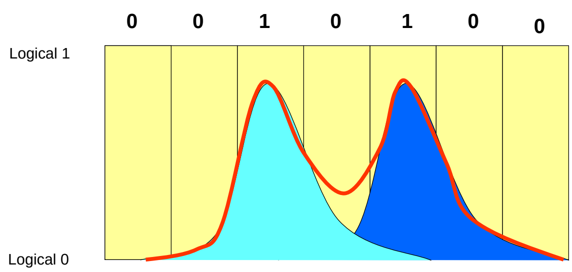
Message: Pre-emphasis allows 3Gb/s on a backplane

Oh really? Just what was that limit again?

The pre-emphasis technique got the job done at the right price
It has limits, the more pre-emphasis, the greater the cross-talk
But now 10Gb switches are common,
- so how do you interconnect them
What's the next speed point
40Gbps (Tracking SONET)
or
100Gbps (Tracking Ethernet *10)
YOUTUBE and similar provides the market
200 terabytes/day
Question:
Where do we go now to get the needed bandwidth?
Answer:
Back to basics

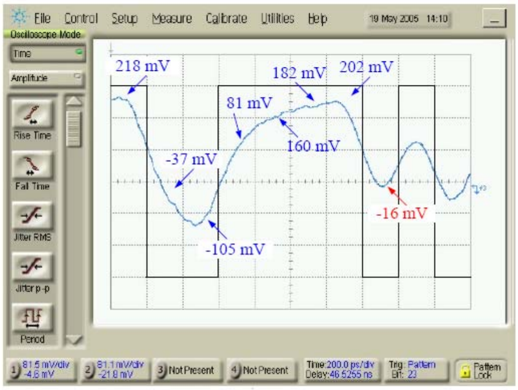


Inter-Symbol Interference

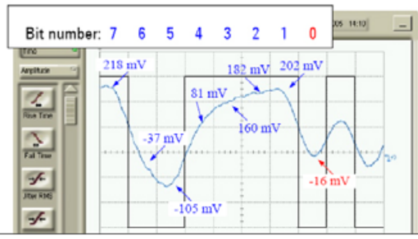


Received signal is the sum of multiple overlapping smeared pulses
For any given channel this sum is deterministic
The signal at any point in time is a function of the preceding bit sequence.

http://cp.literature.agilent.com/litweb/pdf/5989-3777EN.pdf

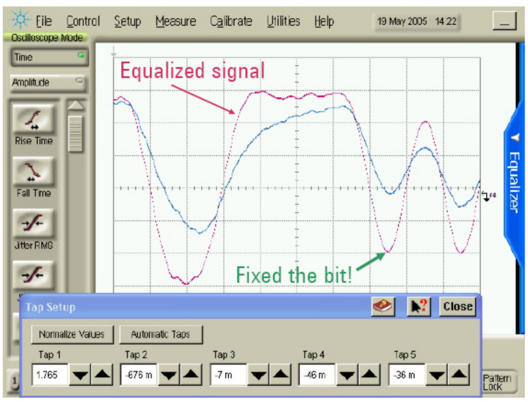


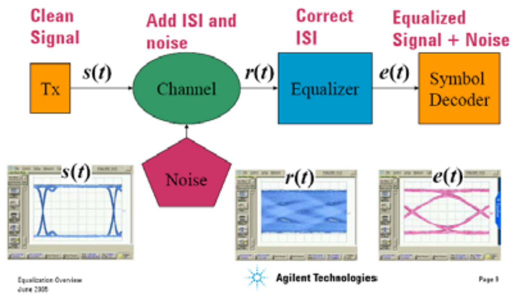
"Tap" the "pre-cursors" to "equalize" the bit



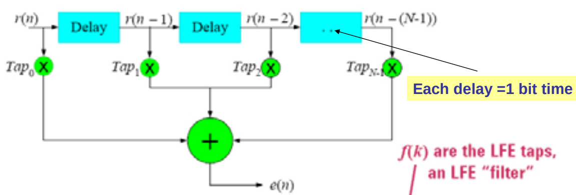
Equalized signal, $e(0) = -16k_0 + 202k_1 + 182k_2 + 160k_3 + 81k_4 - 105k_5 - 37k_6 + 218k_7$

- k_i are correction constants called "Taps"
- The bits before the bit of interest are called "pre-cursors"
- The bits after the bit of interest are called "post-cursors"





LFE: Linear Forward Equalizer

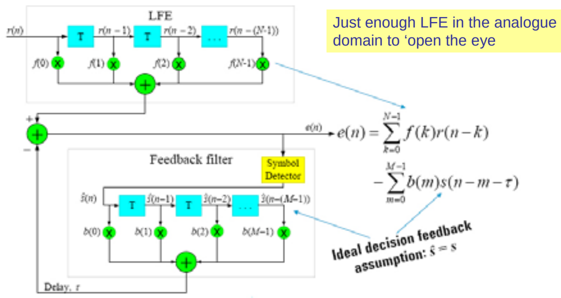


gives $e(n) = \sum_{k=0}^{N-1} Tap_k r(n-k)$ or $\boxed{e(n) = \sum_{k=0}^{N-1} f(k) r(n-k)}$

Unfortunately amplifying some signals also amplifies some noise

.

DFE: Decision feedback Amplifier



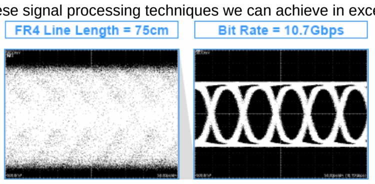
$$e(n) = \sum_{k=0}^{N-1} f(k) r(n-k) - \sum_{m=0}^{M-1} b(m) s(n-m-\tau)$$

Then make a best guess at the symbol then work in the digital
Issues: if you guess wrong you introduce a burst of errors
How do you work out the tap weights
How do you deal with slow changes in transfer function (eg temperature)

For the taps:
- Measure the channel and fix the taps
  however not all channels on one board will be identical
Or
-Use a training sequence in the start-up protocol
Or
-Measure the eye and adaptively feedback information to adjust for a maximum

For slowly changing conditions :
Adaptive techniques.

With these signal processing techniques we can achieve in excess of 10Gbps



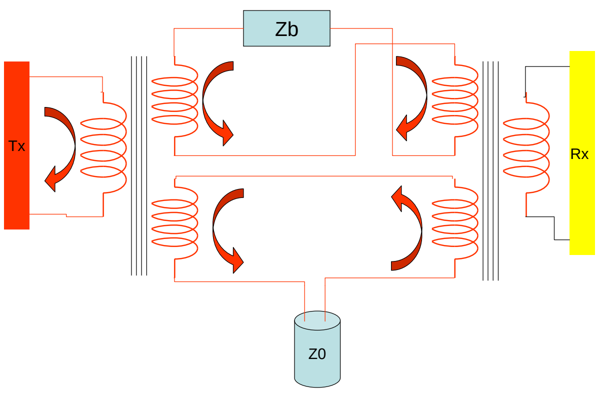
And now for the cable

As we just showed
- 10Gb/s short haul over 8 pairs,(4 in each direction)
- It works very well, probably can push it to 5Gb/s
- But not much more, and this required brute force

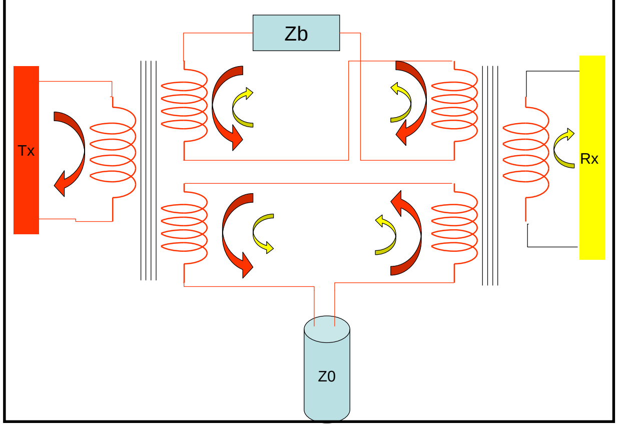
It is just not going to work for 100m of cable
- Need some other way to divide down bits/sec
  to reach an acceptable modulation frequency range
- Encoding can't do it
- Parallel can't do it

Goodbye binary digital, hello multi-level signaling
And there is no room for 8 pairs, hello hybrid coupler

The Hybrid Coupler:1



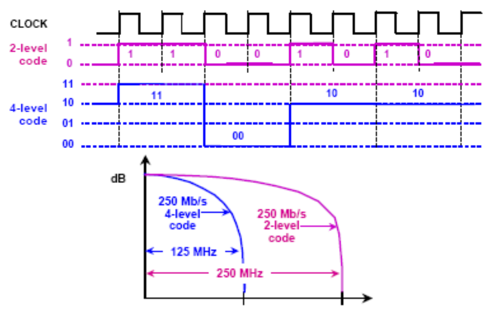
The Hybrid Coupler:2



First we look at 1Gb Ethernet

Divide by four using four channels
250Mb/s over 4 channels=1Gb/s
Encode 2 bits/symbol with redundancy for error
Resulting 125 Mbaud (~80MHz bandwidth):

- transmissible with low interference radiation



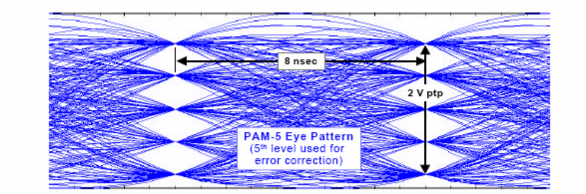
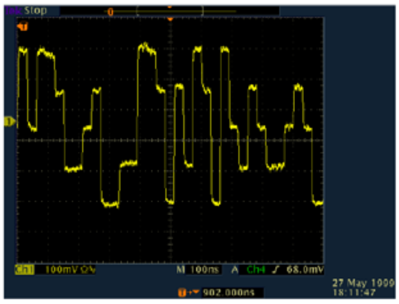


Figure 3: PAM-5 Eye Pattern. 1000Base-T generates a 5-level 2V peak to peak data signal with the symbol period of 8 nsec.

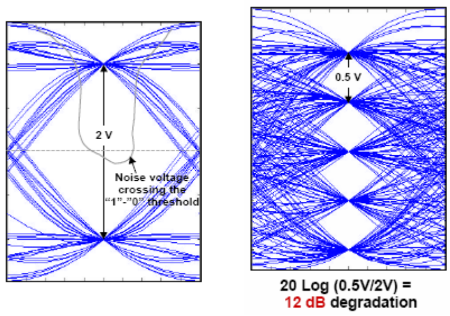
PAM5 Signal Appearance Example



30 MHz
60 MBaud

Still many problems to overcome

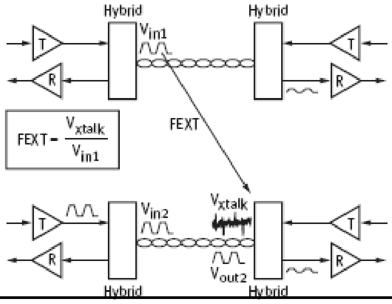
We reduce the symbol rate to fit the bandwidth
But have traded bandwidth for signal to noise problems



Must be much more attentive to self generated noise

FEXT

By using multiple conductors we have added crosstalk problems



ECHOS

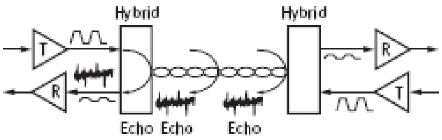


Figure 9. *Echo in Bidirectional Transmission Systems*

NEXT

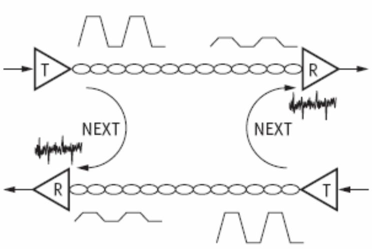


Figure 13. *Near-End Crosstalk (NEXT)*

DELAY SKEW

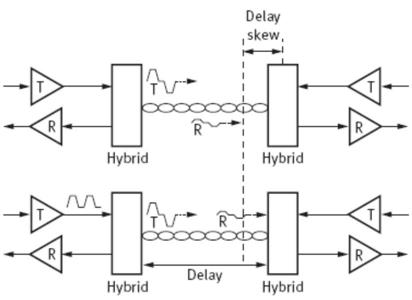
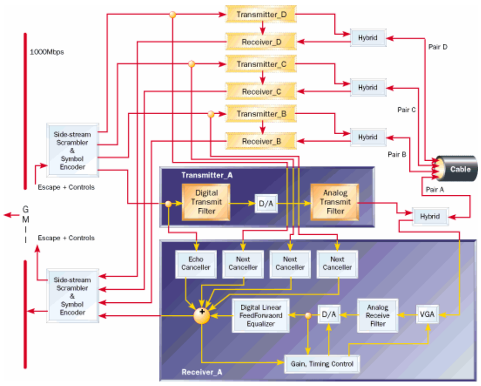


Figure 14. *Delay and Delay Skew*

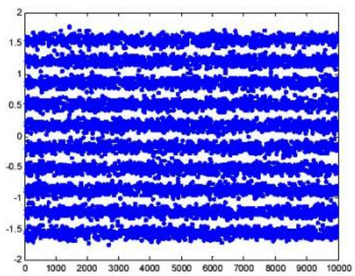


The jump to 10G: More of the same

1000Base-T	10GBase-T
5-level PAM (2b/symbol)	10-level PAM (3 b/symbol)
8-state trellis over 4 pairs	8-state trellis over 4 pairs
Full-duplex echo canceller	Full-duplex echo canceller
125Mbaud, ~80MHz BW	833Mbaud, ~450MHz BW
No FEXT cancellation	FEXT Cancellation required

Throughput = 4pairs x 833Mbaud x 3b/baud = 10Gb/s

No more Digital Electronics only Analogue
DSP experts only need apply



Message

Understand fundamental transmission issues.
Do the best you can with connectivity engineering
Keeping costs to an affordable level
Everything else is signal processing!