candITmedia

vrt medialab

candITmedia

Effective 'Media Services' with 'Efficient IT Technology'

EBU Storage Workshop 21-22 Nov 2011

Media Application - File System - Storage Behaviour

Relevance of Specifications

Luc Andries - CTO CandIT-media *(a VRT-media-lab spinoff)*

luc.andries@candit-media.com

Setting the Scene

Relevance of Specifications

- **Production Media Storage**

(Editing, Transcoding, Rewrapping, ..., ~ File-based Production – ???Media Cloud ???)

- **Guaranteed Predictable Performance**

luc.andries@candit-media.com

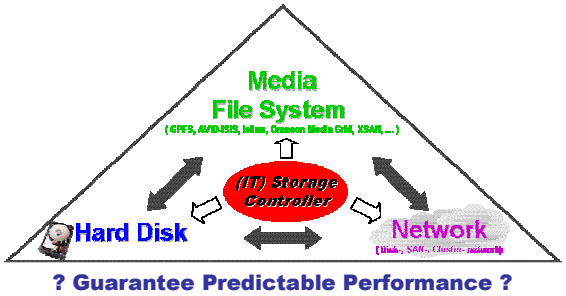
Setting the Scene

Relevance of Specifications

- **Media Application Behaviour**
- **Media File System Behaviour**
- **Media Storage Behaviour**

luc.andries@candit-media.com

(Media) Application Behaviour

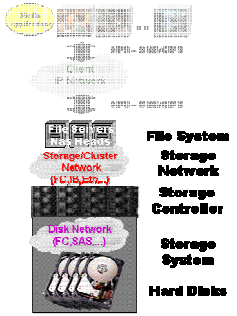


? Guarantee Predictable Performance ?

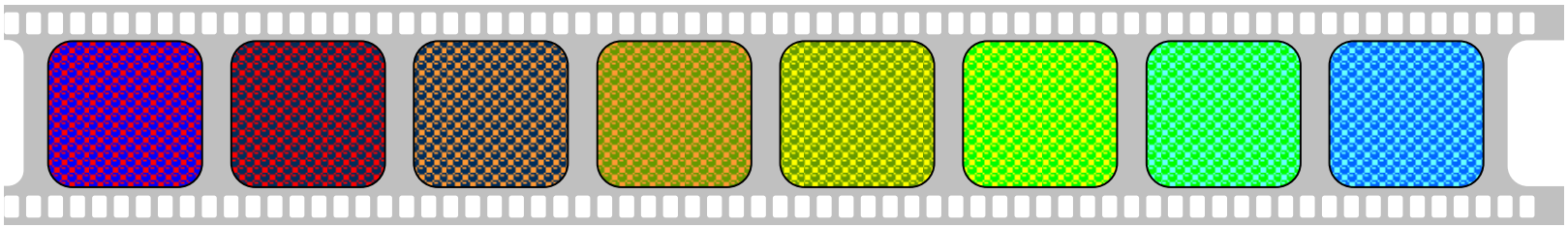
Common View of Media-File Access

~

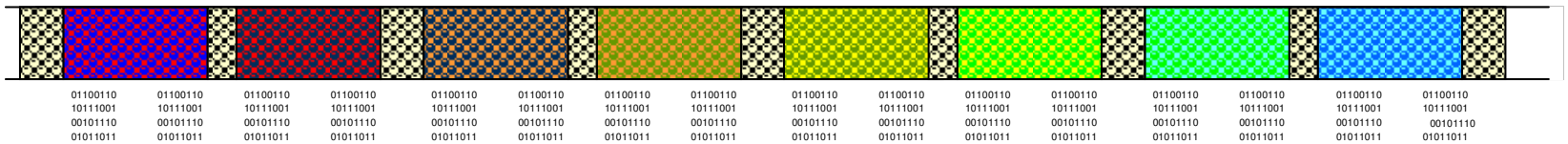
"Sustained Sequential Access" of Large File(s)



'Video'



Media File



Multiple GBytes

Reading/Writing → **= Sequential Access of Bytes**

? How are these Bytes stored on Media Storage ?

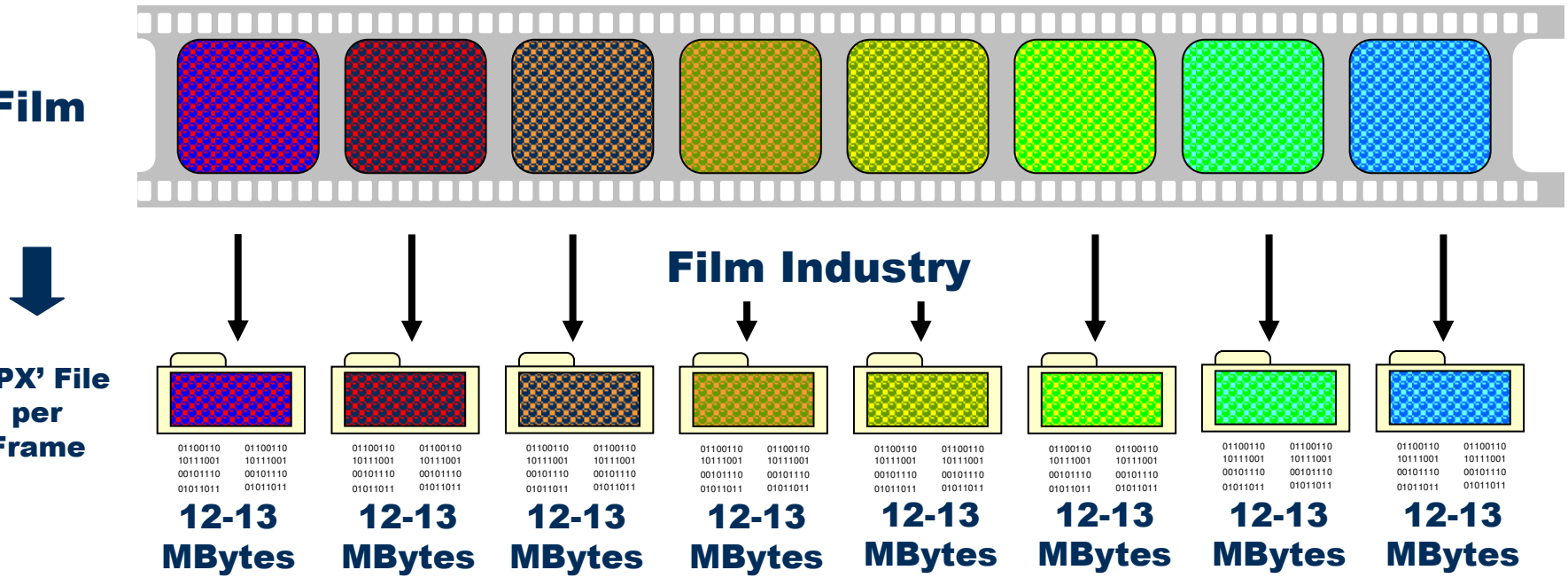
luc.andries@candit-media.com

CONFIDENTIAL

vrt medialab

candITmedia

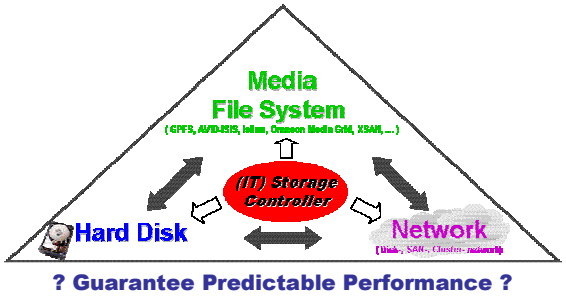
Broadcasting vs Film Industry



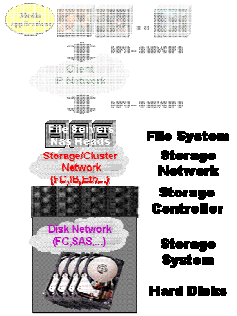
"Sustained *Sequential* Access"
of 'Large' (?) Files

luc.andries@candit-media.com

(Media) Application Behaviour



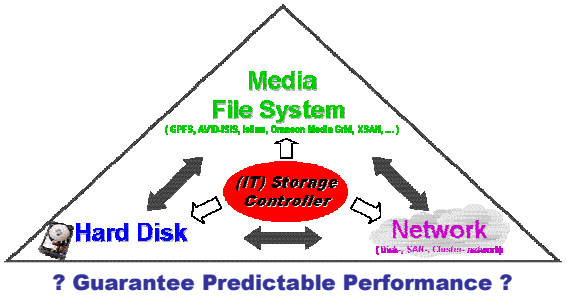
Media Application Behaviour
~



- Access 'Large' Files
- At High Throughput
- In a Multi-User/Multi-Process Context
- At Sustained Load

luc.andries@candit-media.com

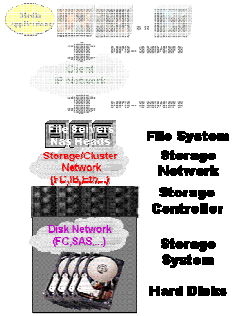
(Media) Storage Specification



? Guarantee Predictable Performance ?

Common Perception of Media-File Access

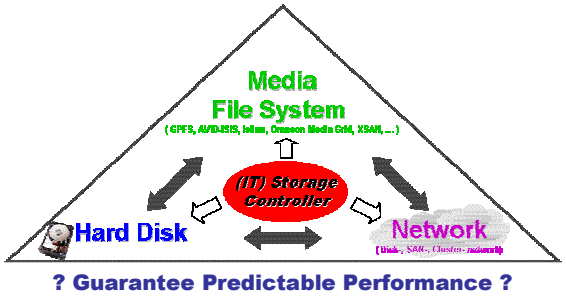
~



"Sustained Sequential Access" of Large File(s)

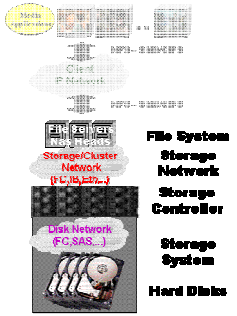
Global Namespace
Aggregated Throughput
RAID 5 (7+1)
N+1 Protection
Stripe size
FC
Parall
Throughput
Failure Groups
NFS
100.000 IOPS
Scale-Up
? How does this
NAS Heads
FCoE
DCB
NAS
MetaData
Fragmentation
SAN
IO from Cache
I/O size
Near Line SAS
Small I/O
ontroller
SMB
Pre-fetch
SAS
(Media) Storage Sp ions
JBOD
Read Ahead
802.3x
Scale-Out
Sequential READS
Mirroring
Block size
SAS II
Guaranteed Throughput
MAID
Large I/O
Sustained Throughput
Rando
R/W
NVRAM
Sequential WRITES
CIFS
SATA
IB
Clustered Filesystem
SSD

(Media) Storage Specification



Common Perception of Media-File Access

~



"Sustained Sequential Access" of Large File(s)

? How does this Translates into ?

(Media) Storage Specifications

Global Namespace
Aggregat Throughput
(7+1)
N+1
Stripe size
Throughput
lure Groups
NFS
100.000 IOPS
Scale-Up
NAS
Small I/O
Fragmentation
SAN
IO from Cache
I/O size
Near Line SAS
SM
JBOD
Sequential READS
Mirroring
Block size
Guaranteed Throughput
MAID
Scale-Out
Large I/O
Sustained Throughput
SATA
NVRAM
R/W
SSD
Sequential WRITES
CIFS
IB
Clustered Filesystem

(Media) Storage Specifications

threaded architecture allows performance to linearly scale with advances in underlying hardware. This same architecture allows the SFA10000 to do what no other RAID controller has been able to do to date: to perform in the extreme range of both throughput and IOPS. The SFA10000 delivers random over 1 million burst IOPS up to and 300,000 random to disk. Sequential block throughput performance is 10GB/s for both reads and writes. Designed to house the most scalable unstructured file data, the system supports up to up to 1,200 drives raw storage while enabling a combination of SAS, SATA or SSD drives.

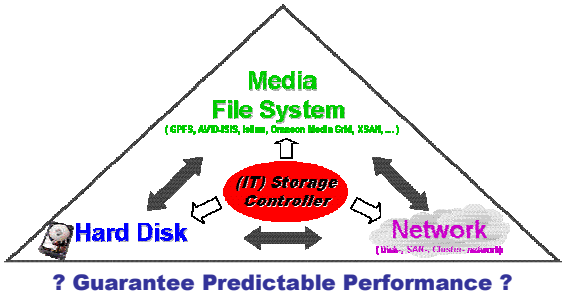
nal primary file storage, providing over 1.4 million NFS ops and 85 gigabyte (GBps) of aggregate throughput, all from a single file system. To accomplish t se extraordinary speeds, the S200 combines the power of ultra-high performance solid (SSD) and 10,000 RPM 2.5-inch Serial Attached SCSI (SAS) drive technology or dual 1GbE and dual 10GbE front-end networking, dual quad core In , a h-performance InfiniBand back-end, and up to 13.8 TB of globally coherent cach dition, the Isilon S200 leverages enterprise SSD technology to accelerate han face-intensive metadata operations, and, also, provides an ability to place the mission ritical, ency-sensitive data on SSDs in a SmartPools™ environment.
celerate your high-concurrent and sequential-throughput applications.

parallel architecture and a hardware accelerate
sustained and predictable performance, innova
performance, quality of service, data protectio
deploy and mange fewer systems to meet today's changing data patterns and
the explosive growth in capacity and performance requirements. Significantly
lowering both acqui os and total cost of ow
ideal for performa , real-time streaming a
to 6GB/s of sustained throughput; and to consolidat
archives, efficiently storing up to 3.6 petabytes in ju

90 % is Irrelevant !

Effortless real time performance and efficient bandwidth delivery set ISIS 7000 apart; With each ISIS 7000 Engine providing up to 400 MB/s of useable bandwidth, predictable, linear performance is sustained even as client counts increase, right up to 4.8 GB/s.

(Media) Application Behaviour



Media Application Behaviour

~

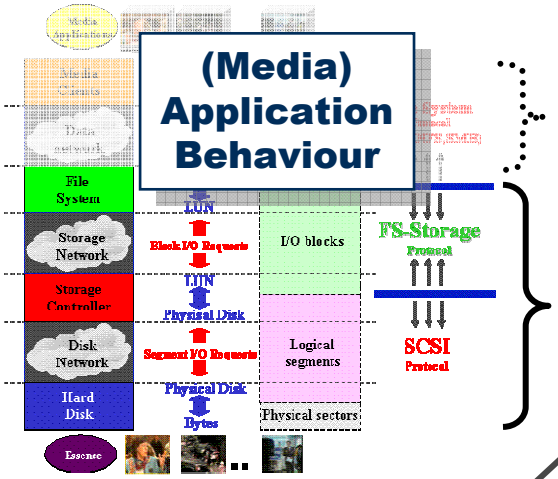
Sequential Access

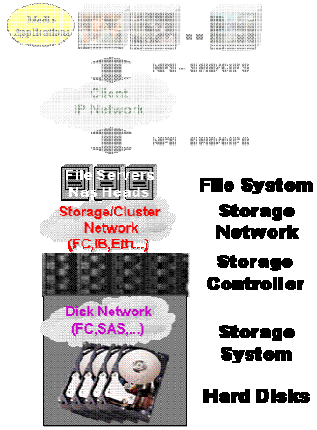
Versus

- Access 'Large' Files
- At High Throughput
- In a Multi-User/Multi-Process Context
- At Sustained Load

luc.andries@candit-media.com

The Basic Triangle





Media
File System
(GPFS, AVID-ISIS, Isilon, Omneon Media Grid, XSAN, …)

? Media

(IT) Storage Controller

? Media

? Media

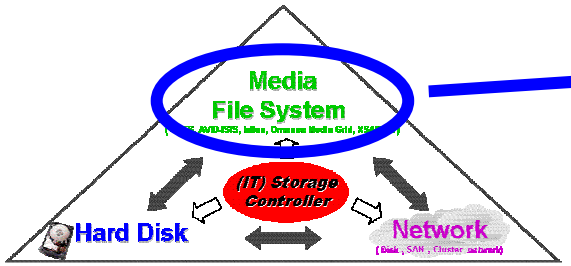
Hard Disk

Network
(Disk-, SAN-, Cluster- network)

? Guarantee Predictable Performance ?

luc.andries@candit-media.com

(Media) File System Behaviour



? Guarantee Predictable Performance ?

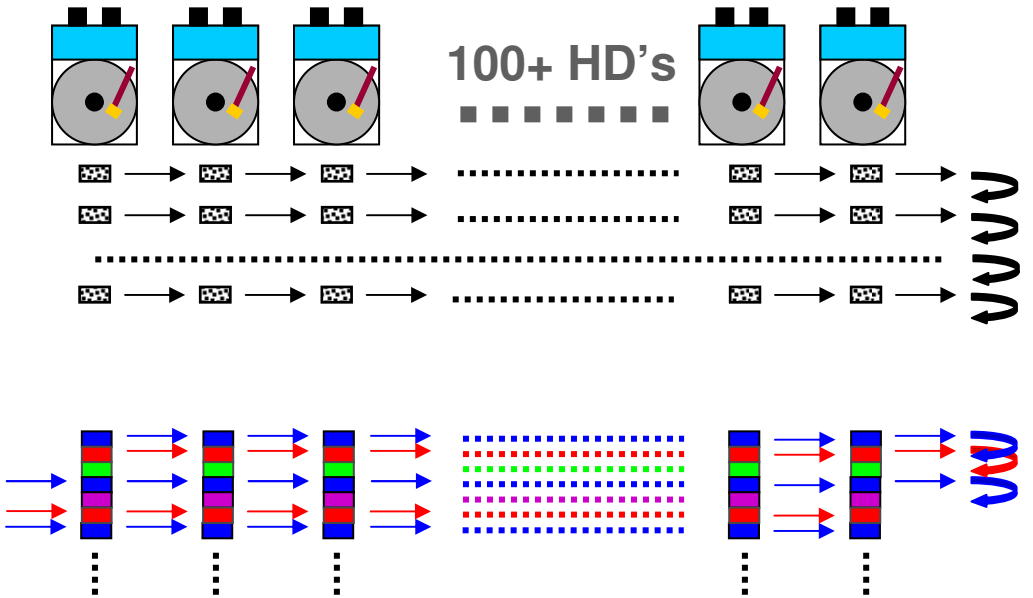
Media File System is only Place for Media Intelligence in Media Storage

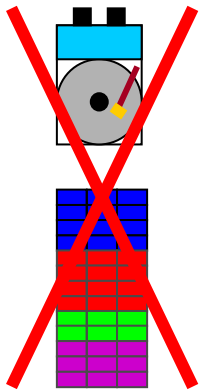
- Access 'Large' Files
- At High Throughput
- In a Multi-User/Multi-Process Context
- At Sustained Load

// Striping:

- Access 'Large' Files
- At High Throughput
- At Sustained Load
- In a Multi-User/Multi-Process Context

Process 1: Media File
Process 2: Media File
Process 3: Media File
Process 4: Media File

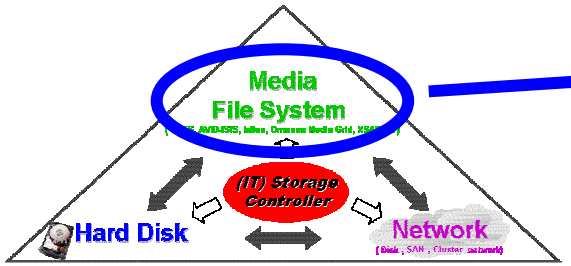




Media File System Write Behaviour Leads to Natural Fragmentation of Disks

luc.andries@candit-media.com

(Media) File System Behaviour



Media File System
Write Behaviour Leads to
Natural Fragmentation of Disks

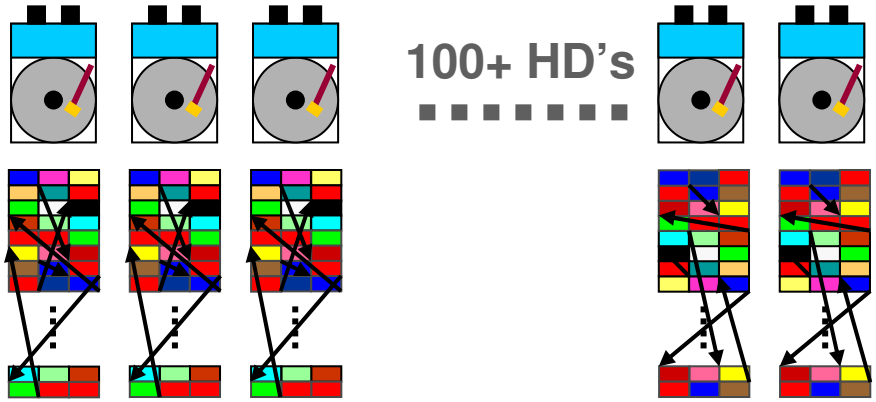
- Access 'Large' Files
- At High Throughput
- In a Multi-User/Multi-Process Context
- At Sustained Load

? Guarantee Predictable Performance ?

// Striping:

Multiple Reading Processes:

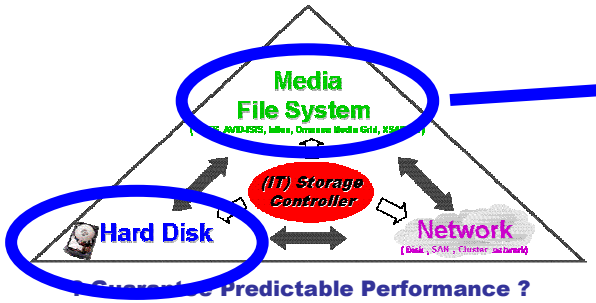
- Process 1: Media File
- Process 2: Media File
- Process 3: Media File
- ⋮
- Process n: Media File



Natural Fragmentation Leads to a
Random Access
Media File System Read Behaviour
on each Individual Disk

luc.andries@candit-media.com

(Media) File System Behaviour



Media File System
Write Behaviour Leads to
Natural Fragmentation of Disks

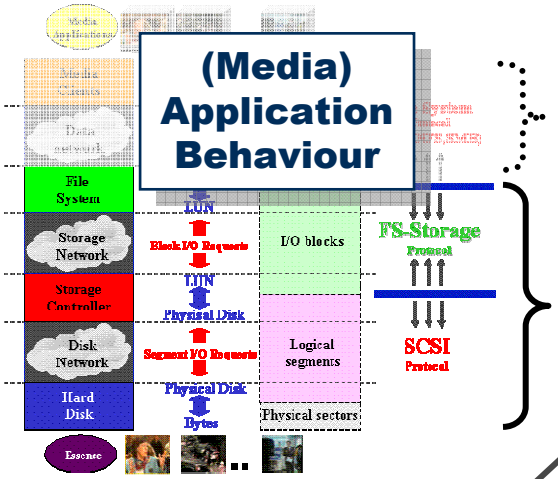
- **Access 'Large' Files**
- **At High Throughput**
- **In a Multi-User/Multi-Process Context**
- **At Sustained Load**

Natural Fragmentation Leads to a
Random Access
Media File System Read Behaviour
on each Individual Disk

Predictable Media Storage Performance
Requires a
Predictable Hard Disk Model
for Random Access

luc.andries@candit-media.com

A Hard Disk Model



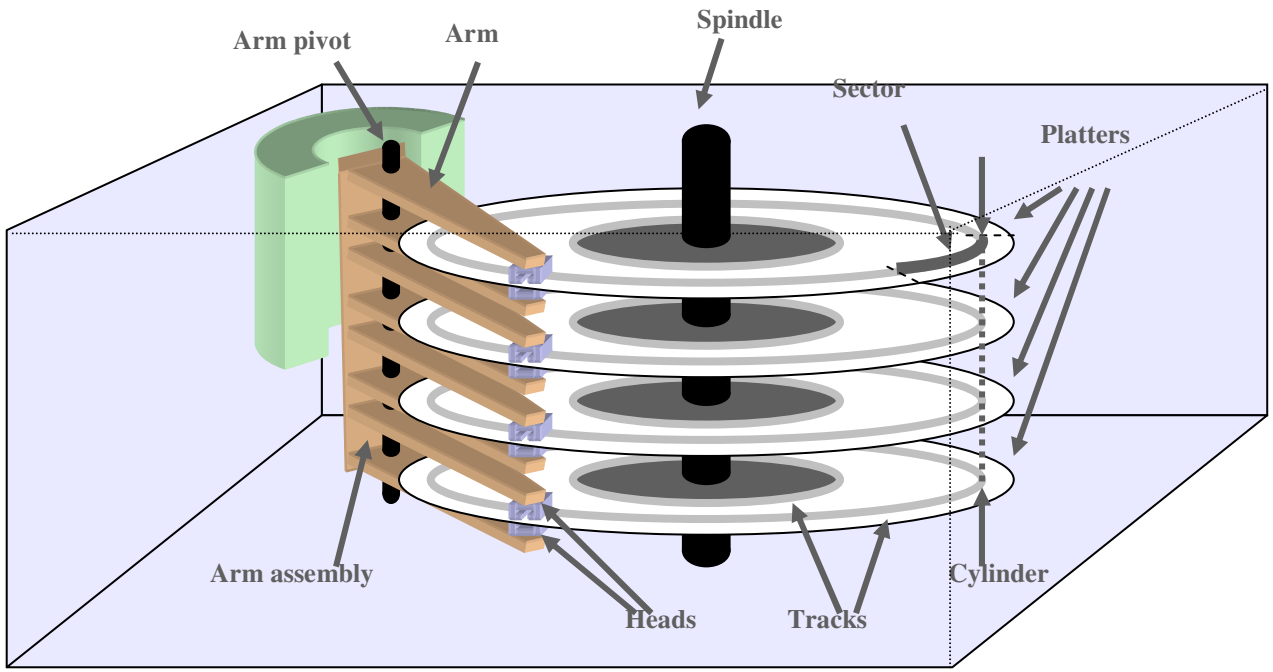
Predictable Hard Disk Model for Random Access

? Guarantee Predictable Performance ?

luc.andries@candit-media.com

A Hard Disk Model

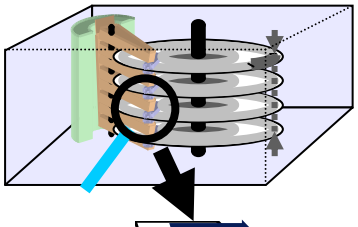
A Physical Hard Disk Model



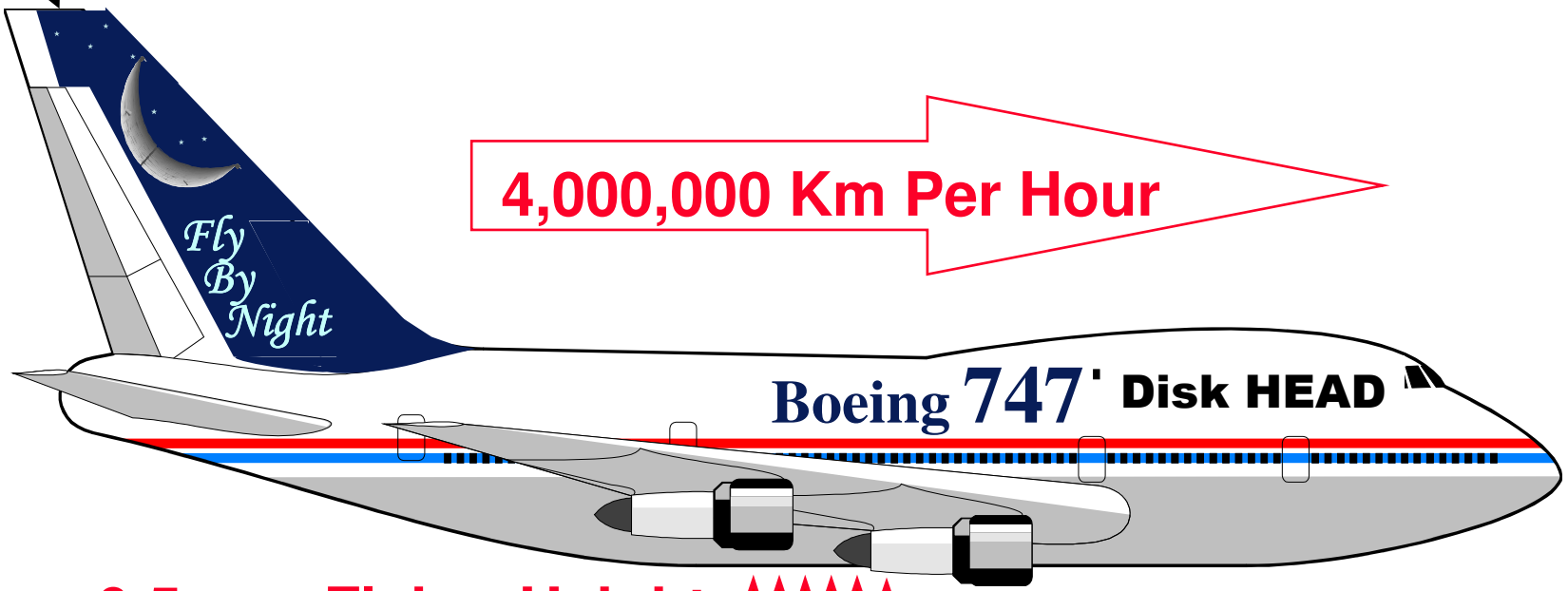
Don't do this at Home !!!

luc.andries@candit-media.com

A 'Physical' Disk Model



Another "Physical" Hard Disk Model



Sideways movement @ 80 g

luc.andries@candit-media.com

CONFIDENTIAL

A Hard Disk Model

SAS vs SATA

Different name comes from Interface Technology

SCSI (Small Computer System Interface)
FC(-AL) (Fibre Channel Arbitrated Loop)
SAS (serial attached SCSI)

⟷

ATA (Advanced Technology Attachment)
IDE (Integrated Disk Electronics)
SATA (serial ATA)

luc.andries@candit-media.com

A Hard Disk Model

SAS vs SATA

Different specifications come from **Target Application Environment**

High End Systems / Reliability	⟷	Low Cost (PC's)
In Groups / Racks / Shelves		Single Disk in PC
Random Access / Performance		Personal Use
Heavy Duty 24h/24h 7d/7d		8h/d 5d/w

luc.andries@candit-media.com

A Hard Disk Model

SAS vs SATA

Different specifications come from **Disk Technology**

2.5" Platter diameter	⟷	3.7" Platter diameter
Spinning speed 15 Krpm		Spinning speed 7.2 Krpm
Bearings at both shaft ends		Bearings at single end
Separate processors (Servo - I/O)		Single processor
Stiffness / Mechanical rigidity		Less robust
Stronger actuators / Motors		
Filters / O-ring sealings / Dessicants		

luc.andries@candit-media.com

A Hard Disk Model

SAS vs SATA

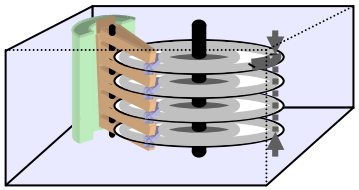
Differences result in **different**

- Data Rate
- Reliability MTBF
- Capacity
- Data Density
- Applications Environment
- Random Access Performance
- Cost

luc.andries@candit-media.com

A 'Physical' Disk Model

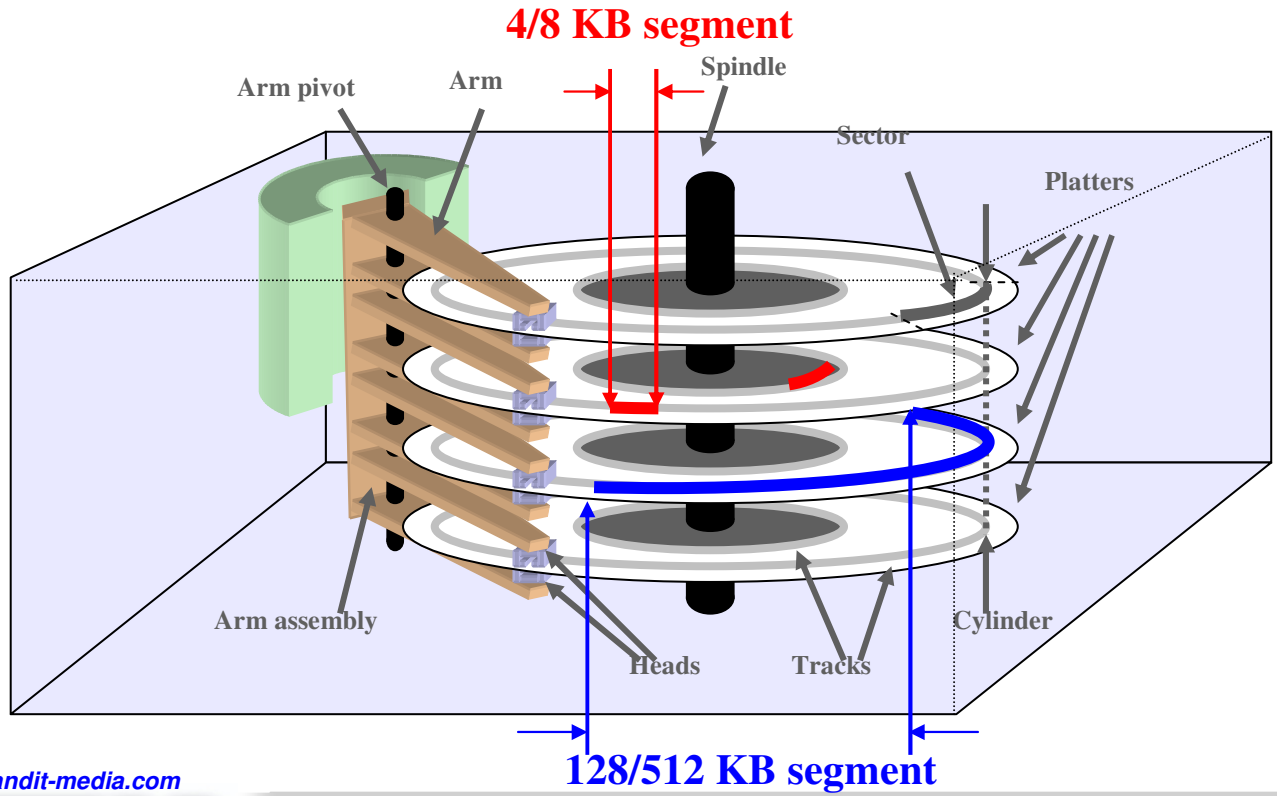
Relevant Parameters



Parameters:
- Lateral seek time
- Rotational speed

} = Head positioning

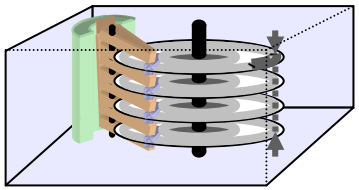
- **Segment size** → = Reading/Writing time



luc.andries@candit-media.com

A 'Media' Hard Disk Model

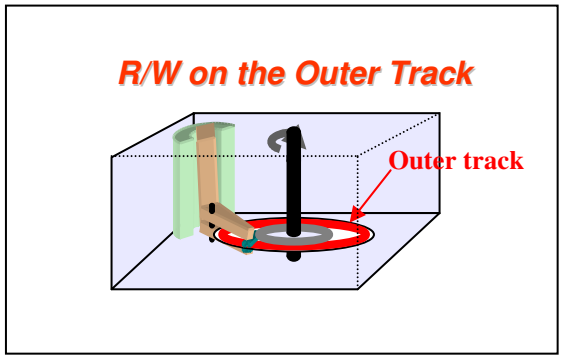
Random Access Behaviour

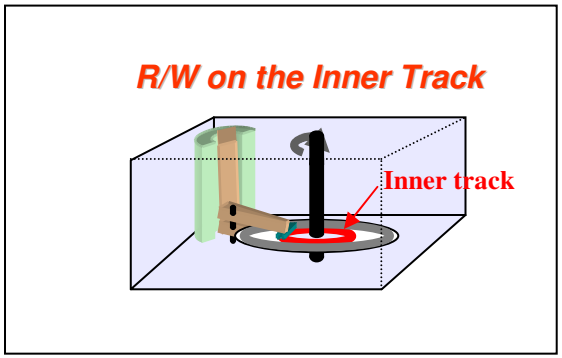


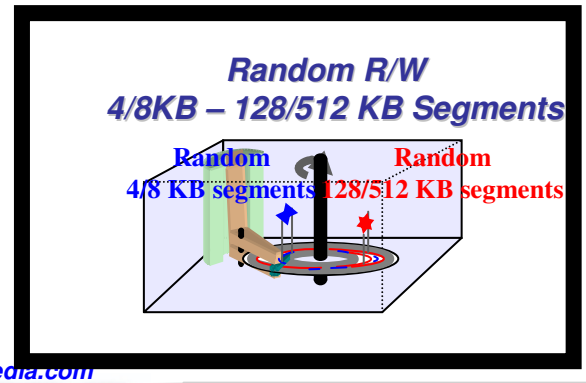
Parameters:
- Lateral seek time
- Rotational speed

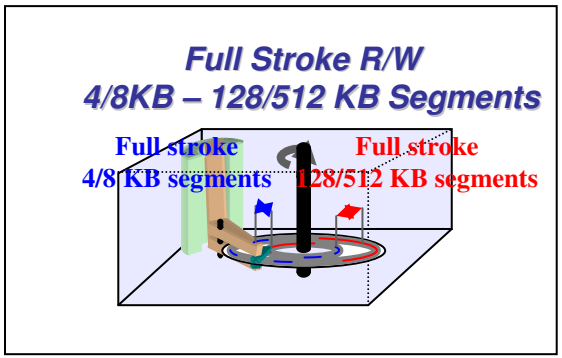
} = Head positioning

- **Segment size** → = Reading/Writing time



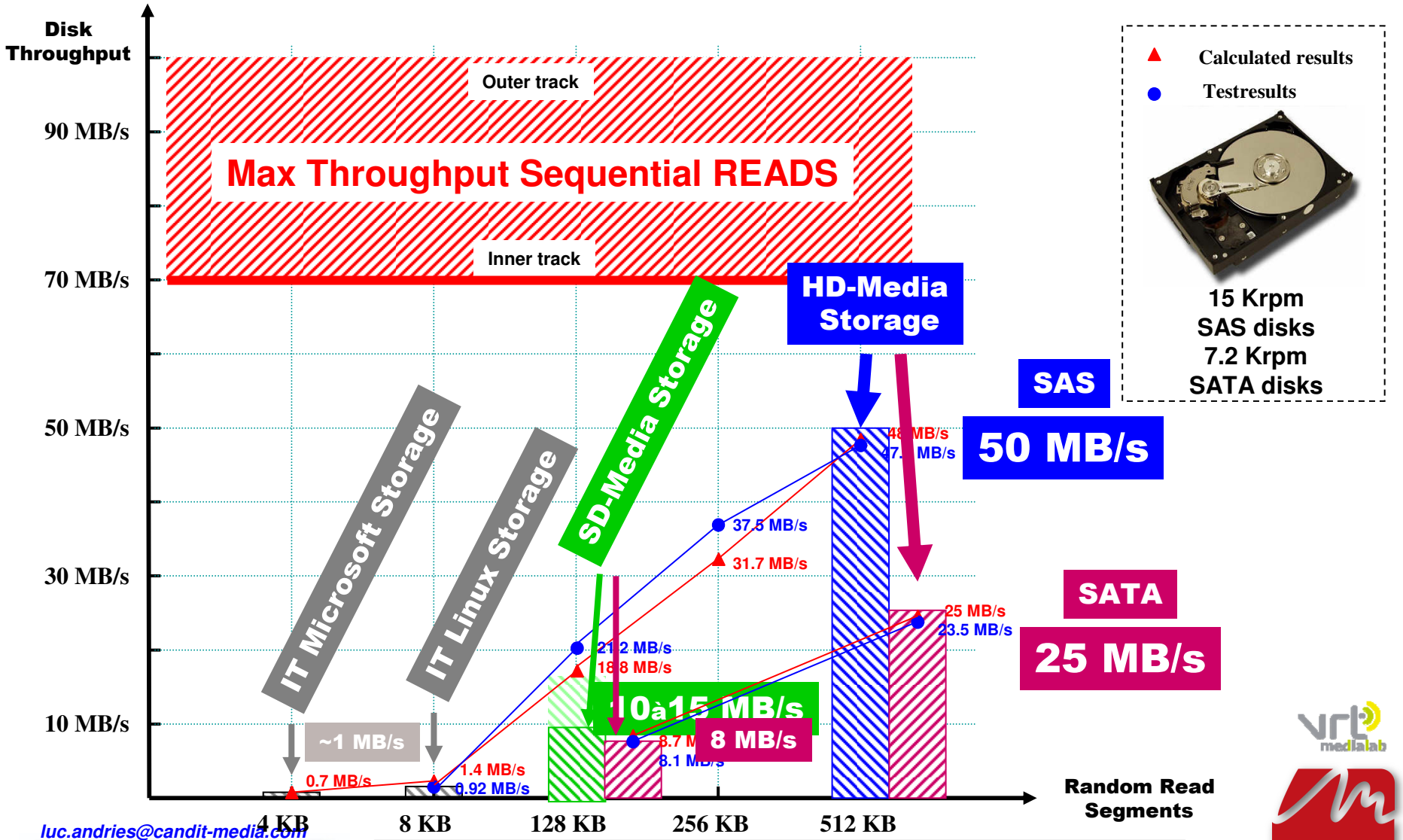






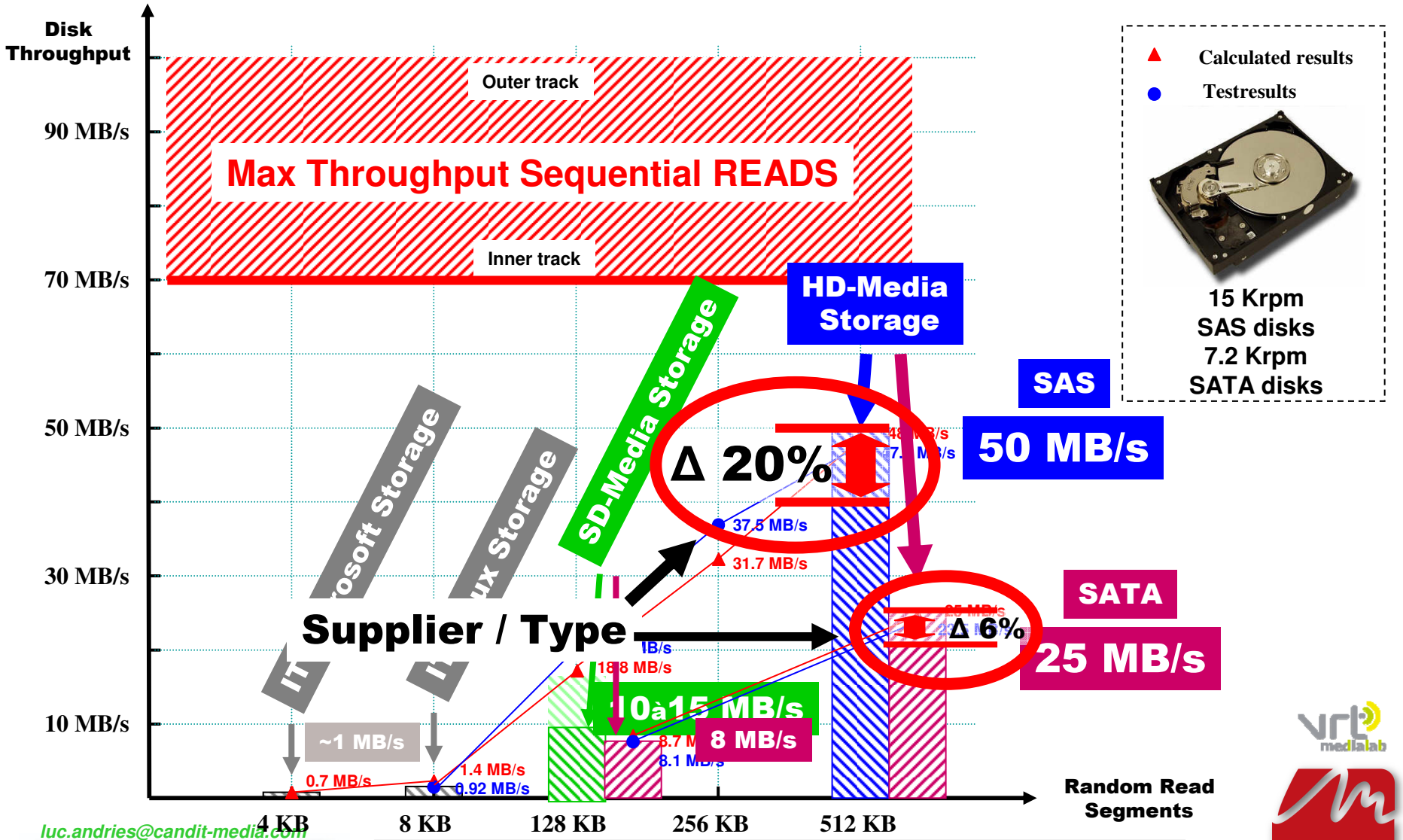
luc.andries@candit-media.com

Some Hard Figures



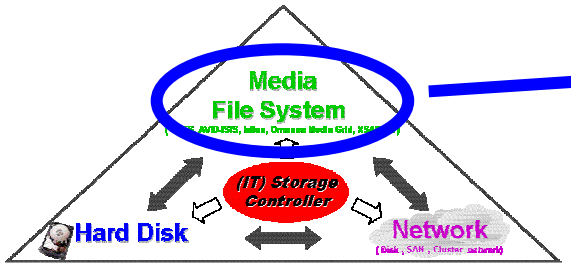
luc.andries@candit-media.com

Some Remarkable Figures



luc.andries@candit-media.com

(Media) File System Behaviour



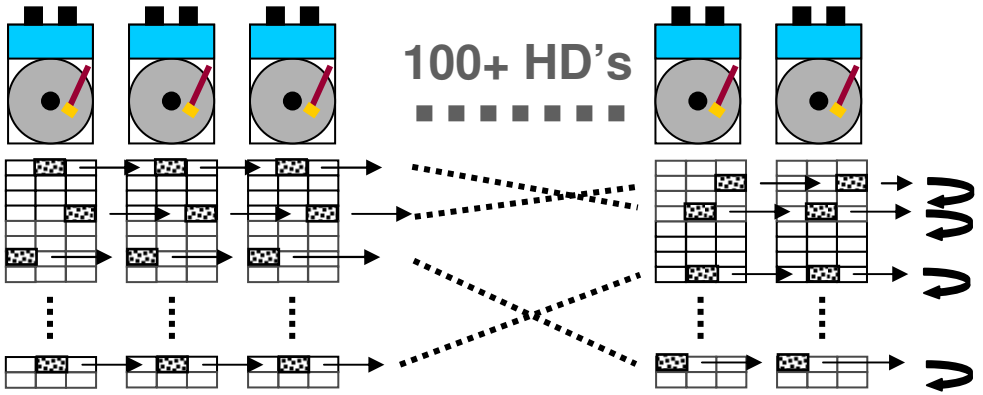
Media File System should Force Scattered Write Behaviour

- Access 'Large' Files
- At High Throughput
- In a Multi-User/Multi-Process Context
- At Sustained Load

? Guarantee Predictable Performance ?

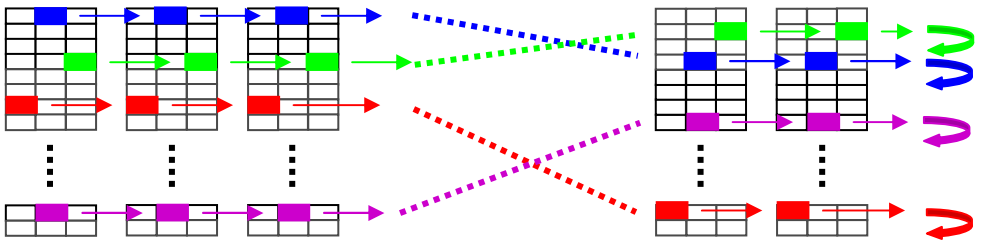
// Striping:

- Access 'Large' Files
- At High Throughput
- At Sustained Load
- Force scattered Writing Pattern



- In a Multi-User/Multi-Process Context

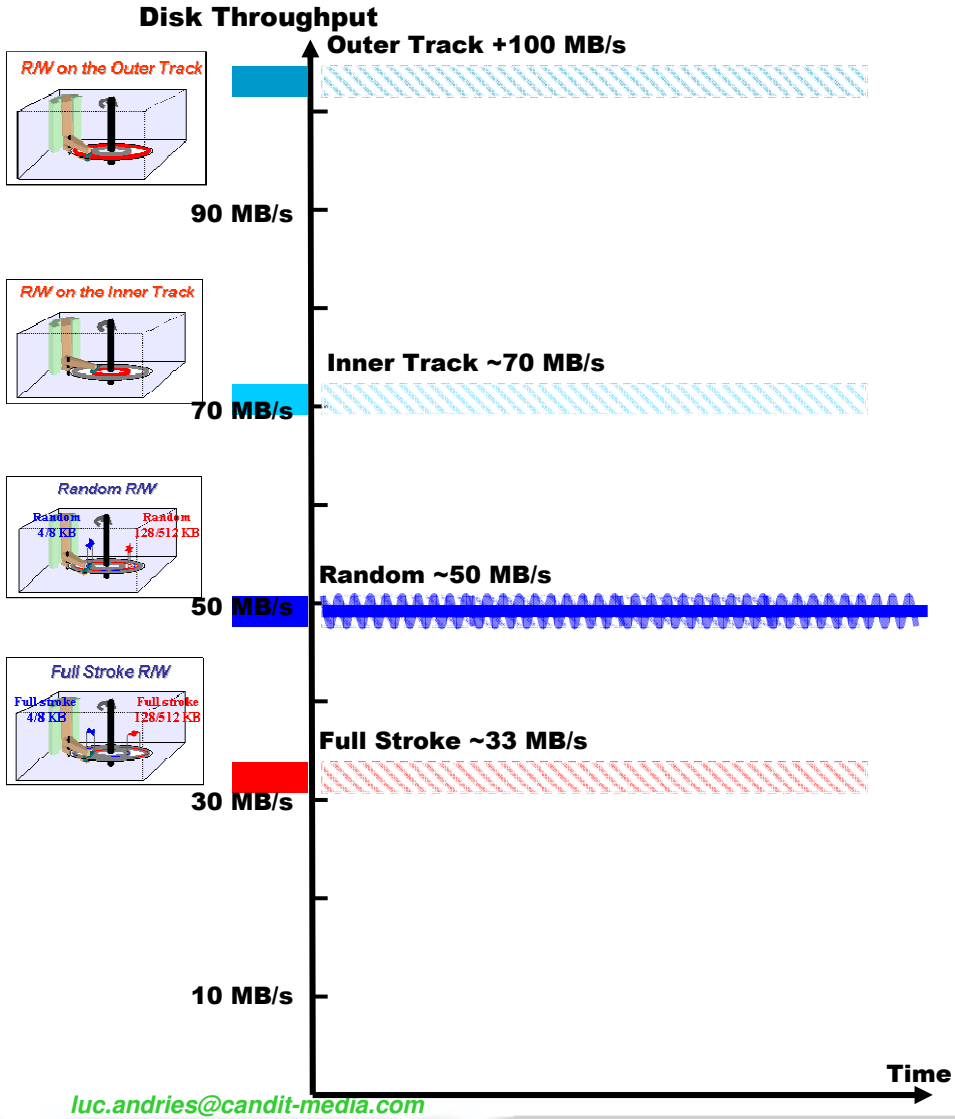
Process 1: Media File
Process 2: Media File
Process 3: Media File
Process 4: Media File

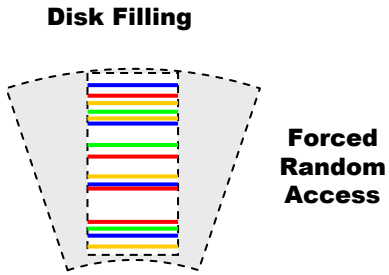


Media File System has to Force Scattered Writing Pattern to get Predictable Disk Reading/Writing Behaviour

luc.andries@candit-media.com

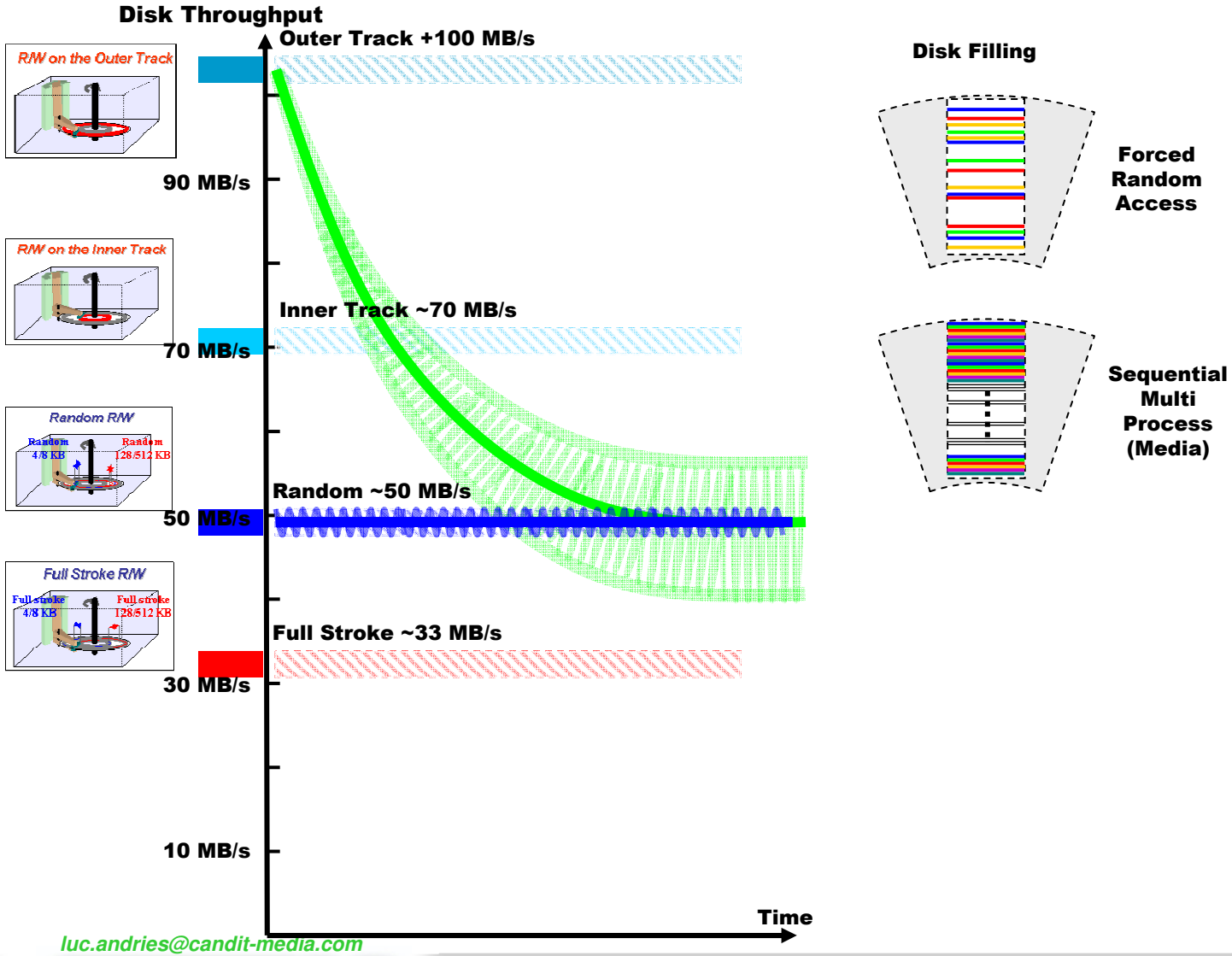
? Random Access = Worst Case ?





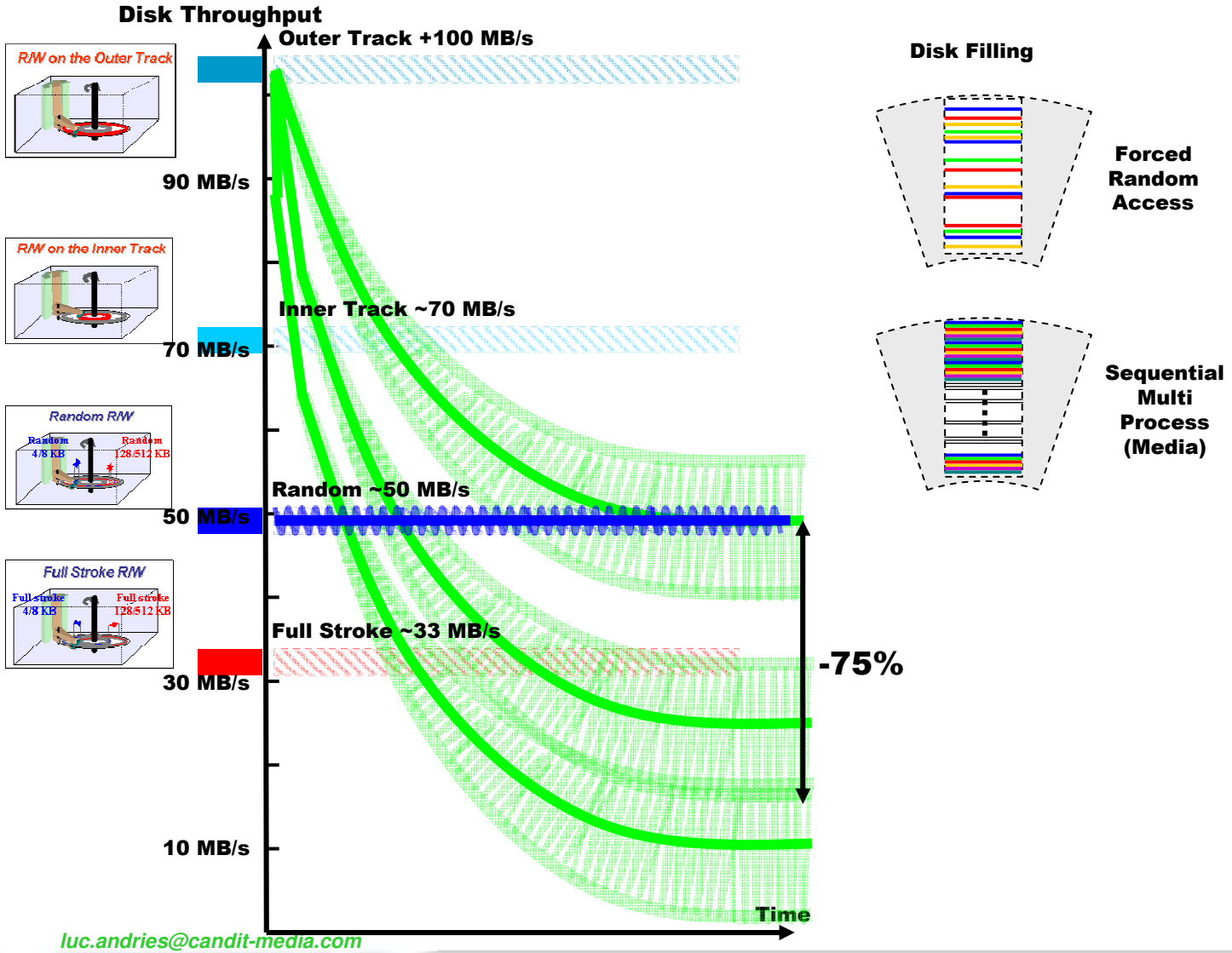
luc.andries@candit-media.com

? Random Access = Worst Case ?



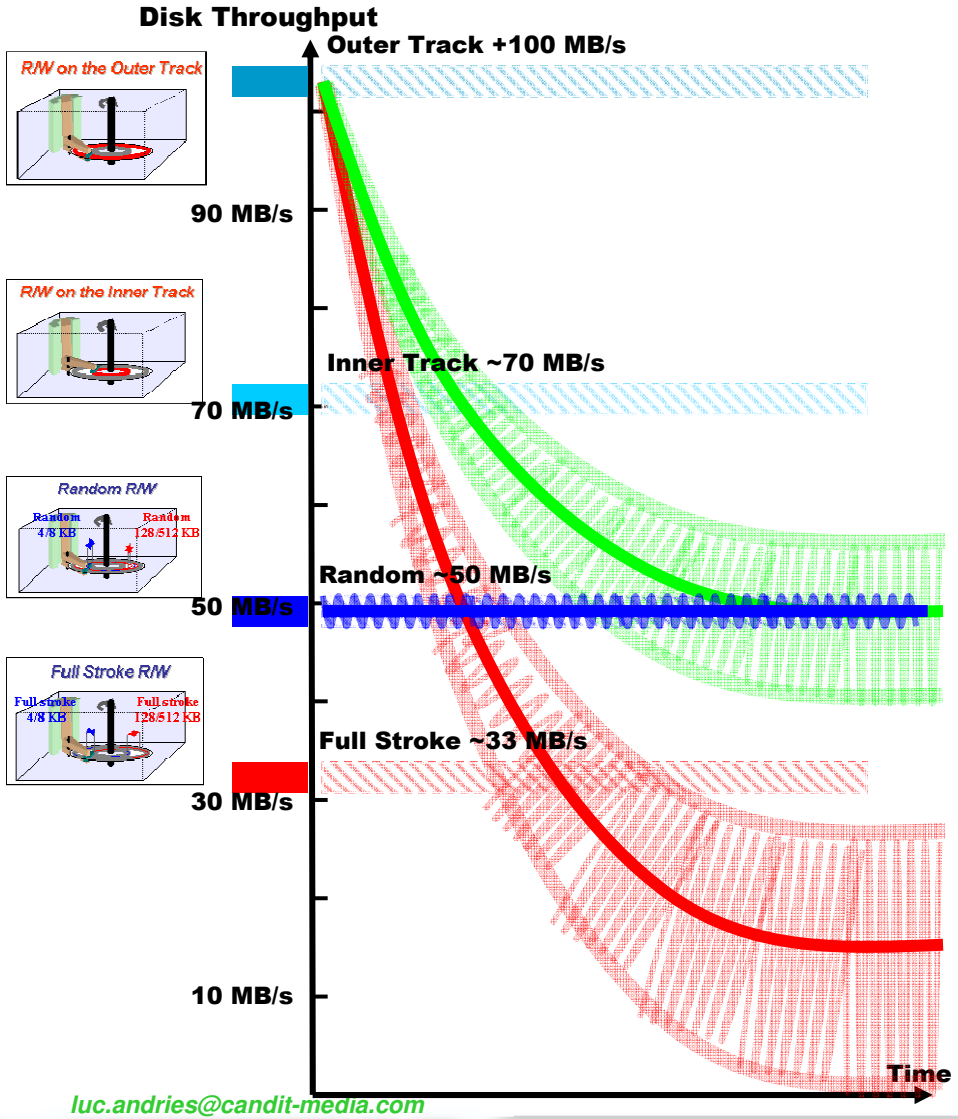
luc.andries@candit-media.com

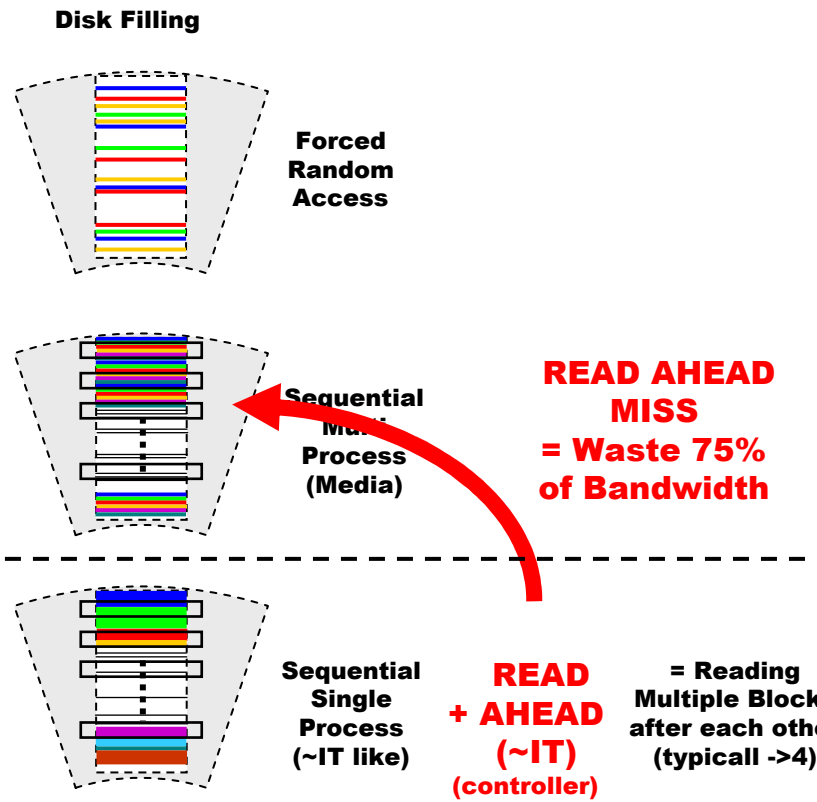
? Random Access = Worst Case ?



luc.andries@candit-media.com

? Random Access = Worst Case ?

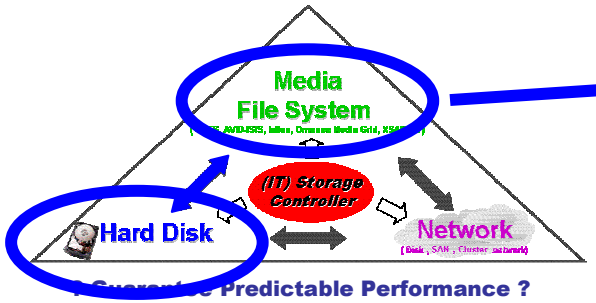




luc.andries@candit-media.com

CONFIDENTIAL

(Media) File System Behaviour



Media File System
should Force Scattered
Write Behaviour

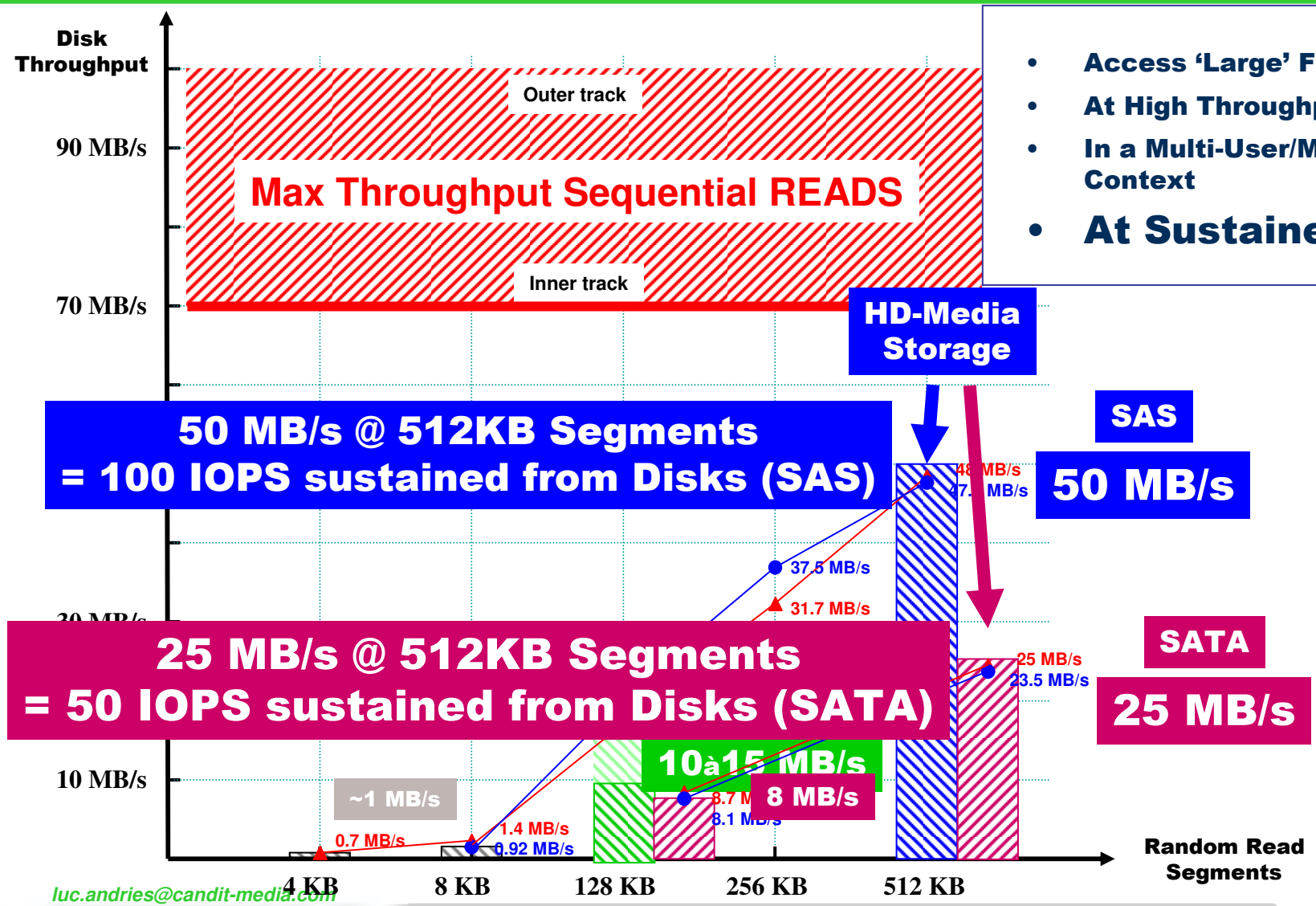
- **Access 'Large' Files**
- **At High Throughput**
- **In a Multi-User/Multi-Process Context**
- **At Sustained Load**

Forced Fragmentation Guarantees
Consistent Random Access
Media File System Behaviour
on each Individual Disk

Guarantees
Predictable
Hard Disk Behaviour

luc.andries@candit-media.com

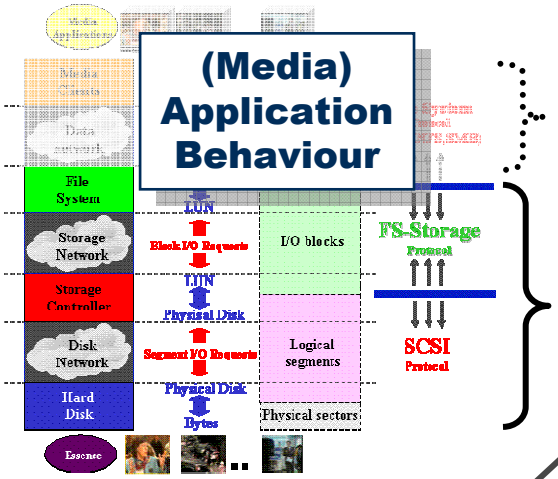
Relevance of IOPS (I/O per Second)



- Access 'Large' Files
- At High Throughput
- In a Multi-User/Multi-Process Context
- **At Sustained Load**

luc.andries@candit-media.com

Media Storage Network



Perfectly Linear Scalable Network between Media File System and Hard Disks

Media File System

(GPFS, AVID-ISIS, Isilon, Omneon Media Grid, XSAN, …)

(IT) Storage Controller

Hard Disk

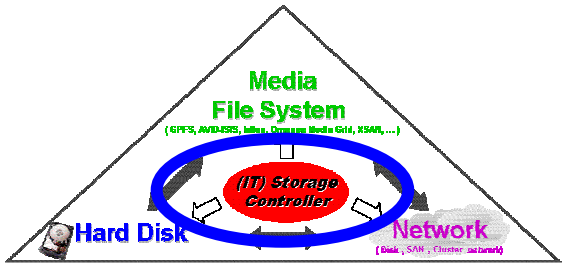
Network

(Disk-, SAN-, Cluster- network)

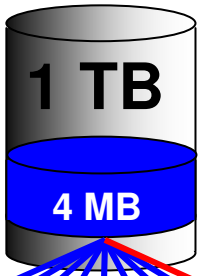
? Guarantee Predictable Performance ?

luc.andries@candit-media.com

'Media' Storage Controller



LUN
(Logical Unit)



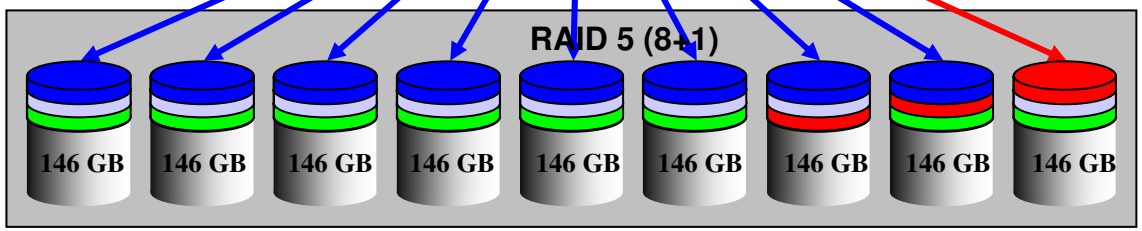
4 MB
I/O block

RAID 5 (8+1) protection

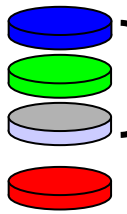
8x512 KB
Disk Segments

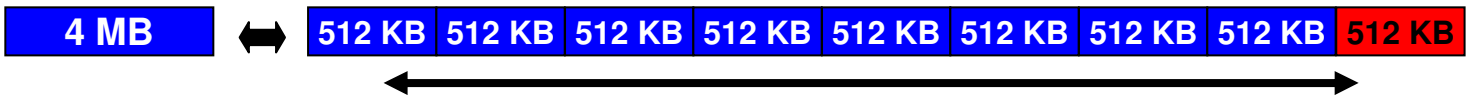
Large stripe depth (4MB) !!!

Disk efficiency => Large segment size
(512 KB)



Physical disks

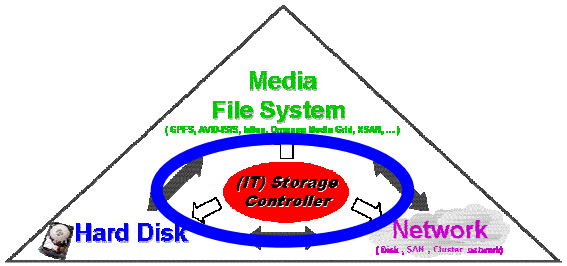




1 I/O (4MB) striped over 8 disks (+ parity)

luc.andries@candit-media.com

'Media' Storage Controller



4 MB
I/O bl

LUN
(Logical Unit)

1 TB
4 MB

RAID 5 (8+1)
prot

8x512 KB

Disk
Segments

Large stripe depth
(4MB) !!!

Disk efficiency
=> Large
segmen

!!! RAID5 N+1 !!! "N" = Power of 2
(!!! RAID6 N+2 !!! "N" = Power of 2)

(8+1)

146 GB | 146 GB | 146 GB | 146 GB | 146 GB | 146 GB | 146 GB

Physical disks

Data

Parity

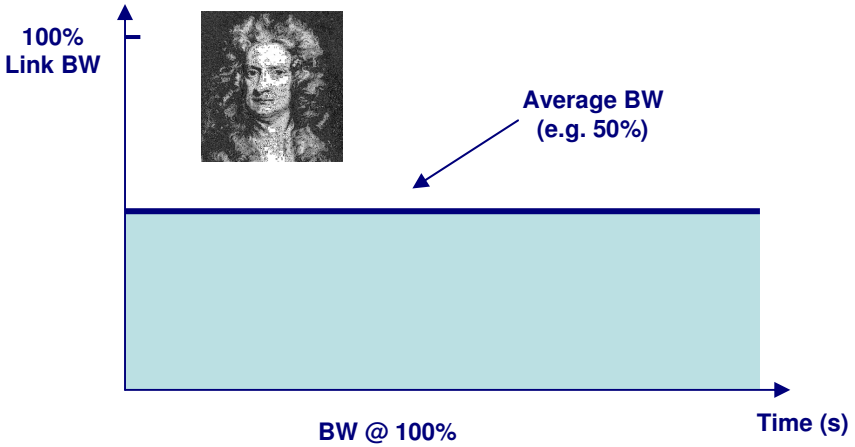
512 KB

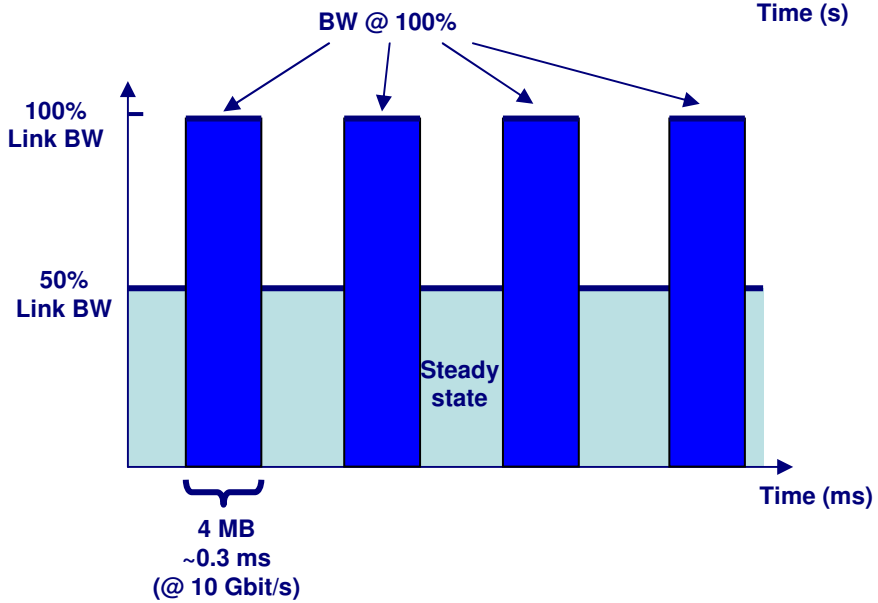
4 | 512 KB | 512 KB | 512 KB | 512 KB | 512 KB | 512 KB | 512 KB | 512 KB | 512 KB

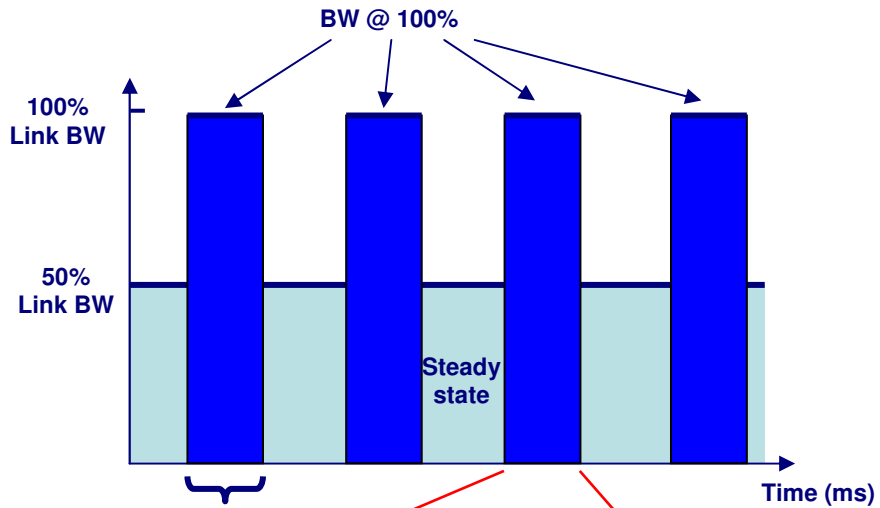
1 I/O (4MB) striped over 8 disks (+ parity)

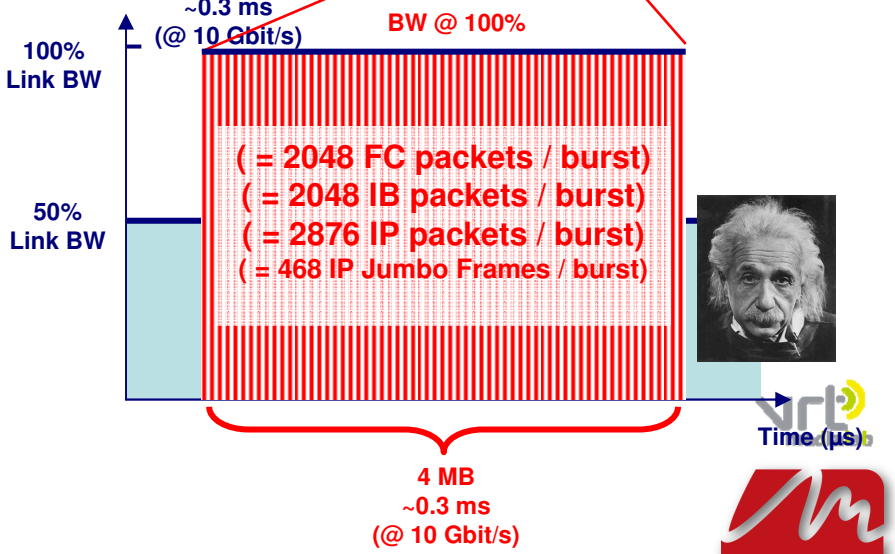
luc.andries@candit-media.com

'Media' Storage Traffic Behaviour



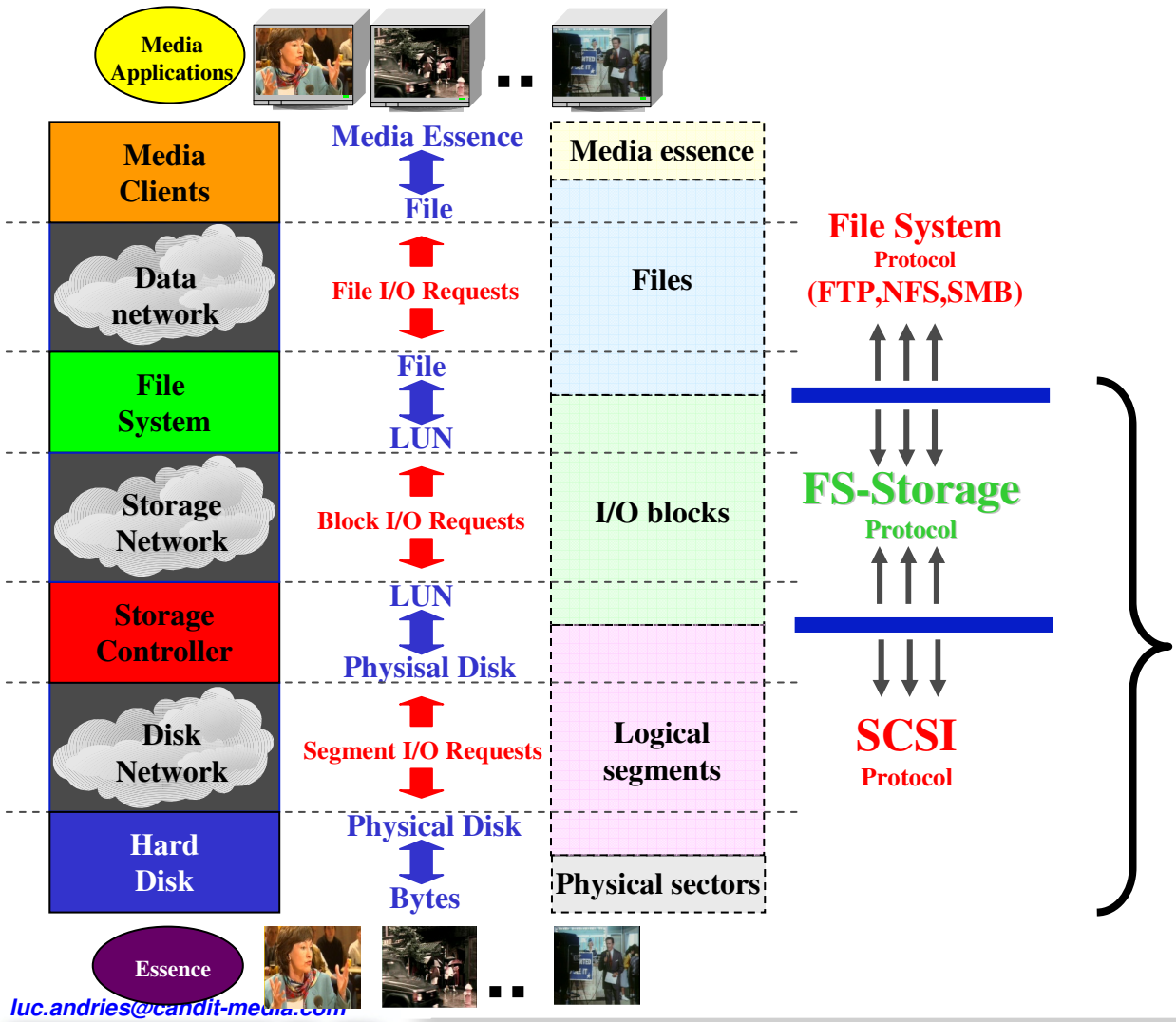






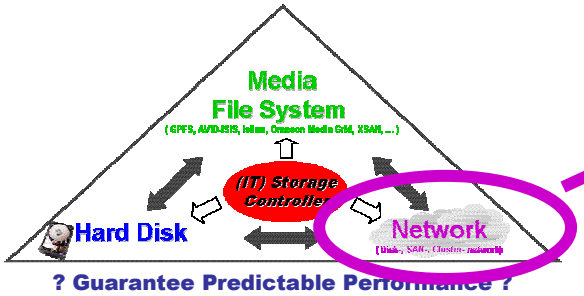
luc.andries@candit-media.com

Media Storage Network

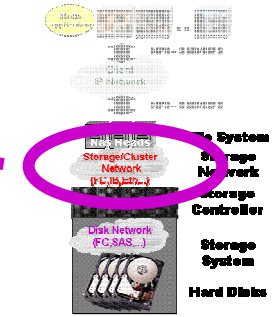


luc.andries@candit-media.com

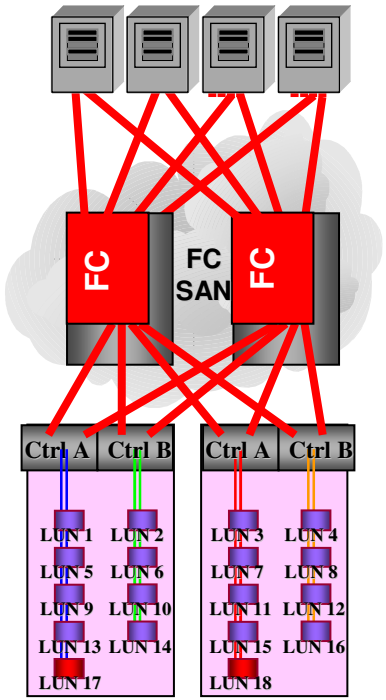
Media Storage Network



Storage Networks
Require
Lossless
Network Technology



Lossless
Technology
by Design

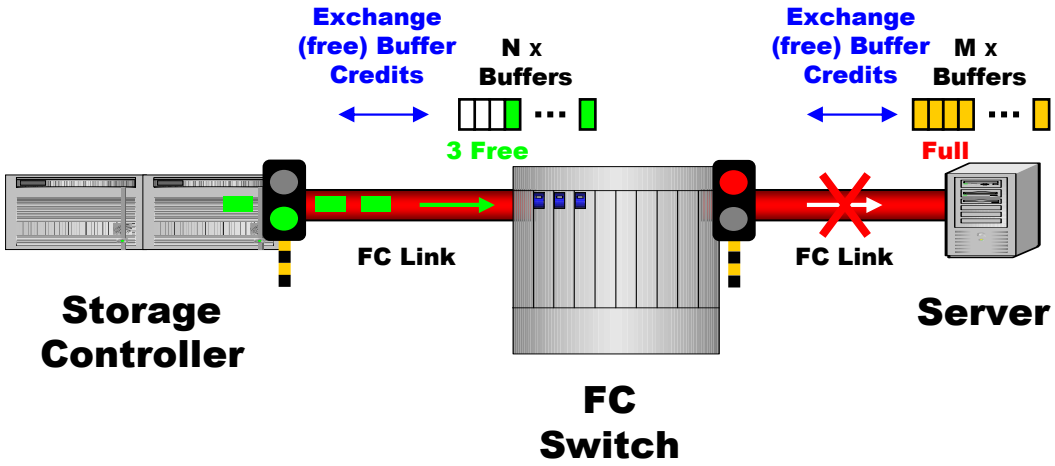


Fibre Channel (FC)
State of the Art
(IT) Storage
Network

luc.andries@candit-media.com

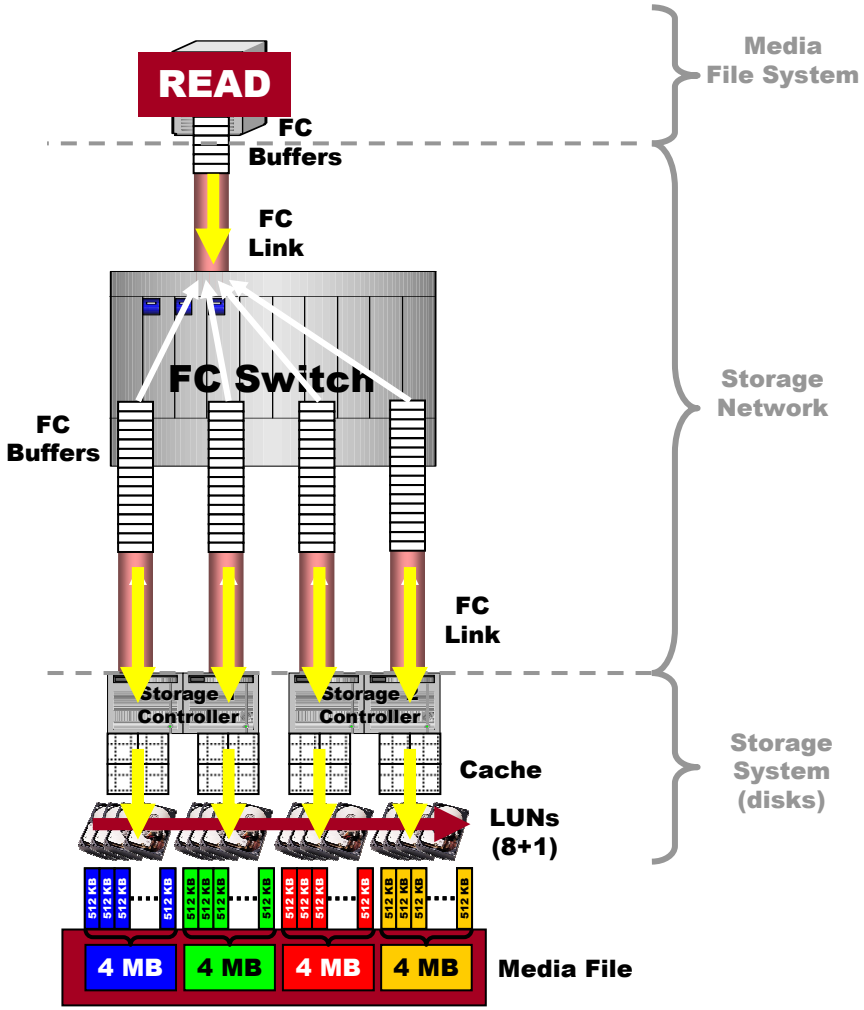
Lossless Fibre Channel (FC)

Lossless by Exchange of Available Buffers (Buffer_to_Buffer Credits)



Packets are only Transmitted if Free Buffers are Available at the (Link)-Destination

FC Storage Network in Media

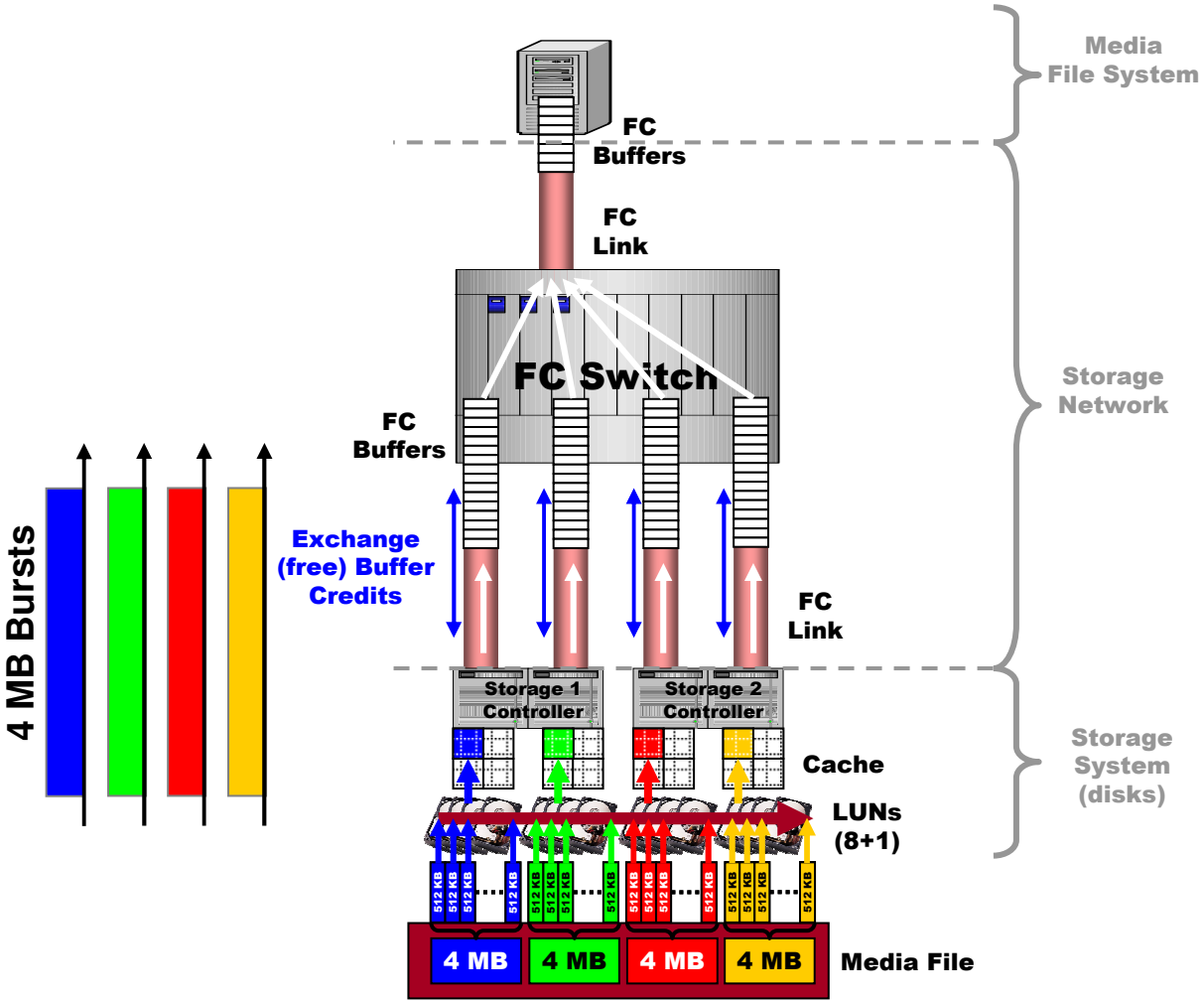


Reading a Media File

- Media File Striping
- Large Block Size/Disk
- RAID 5 (8+1) -> 4MB
- File System READ

luc.andries@candit-media.com

FC Storage Network in Media

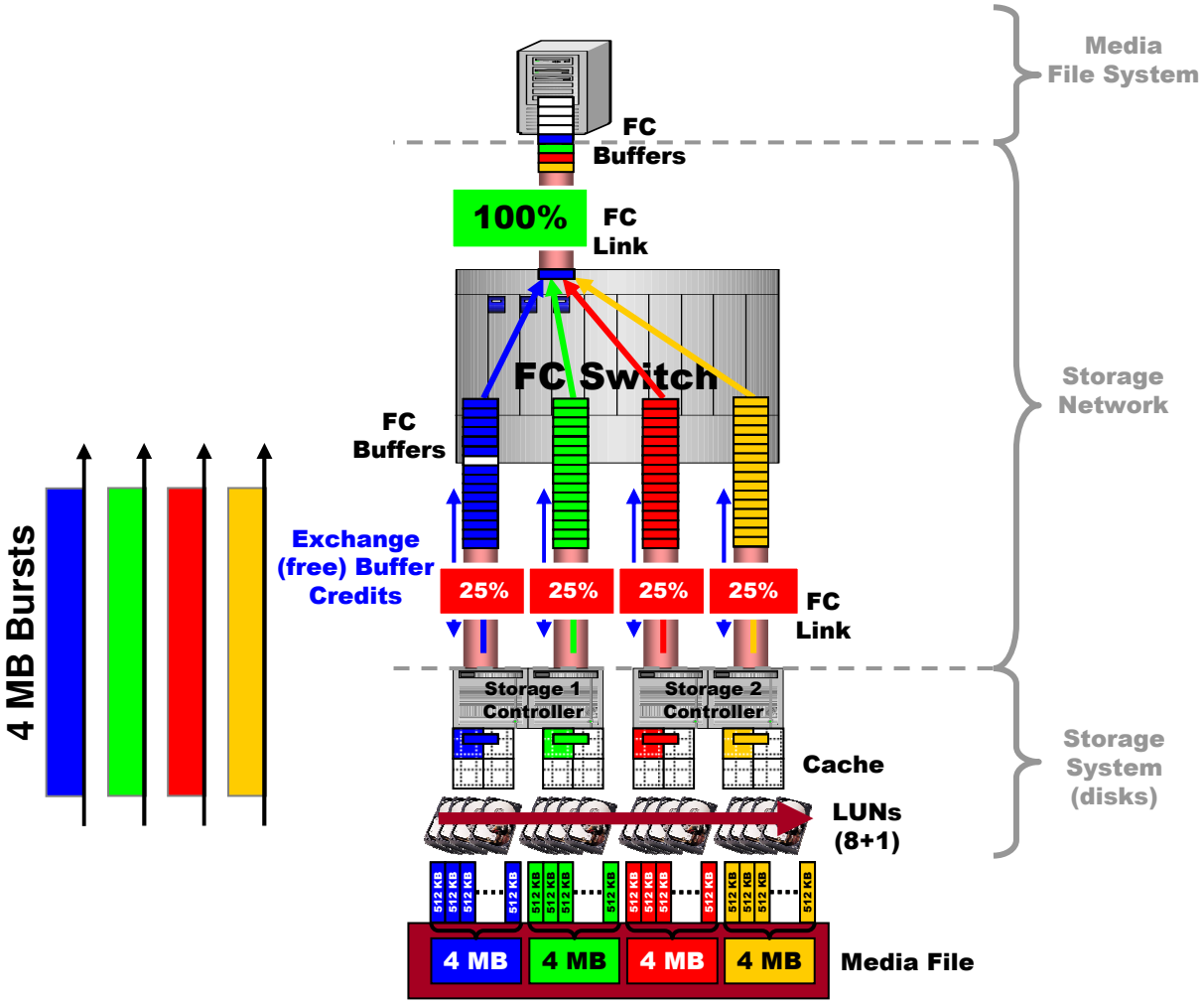


Reading a Media File

- Media File Striping
- Large Block Size/Disk
- RAID 5 (8+1) -> 4MB
- File System READ
- READ from Disk -> Cache
- 4 MB Traffic Bursts
- Buffer_to_Buffer Credits

luc.andries@candit-media.com

FC Storage Network in Media

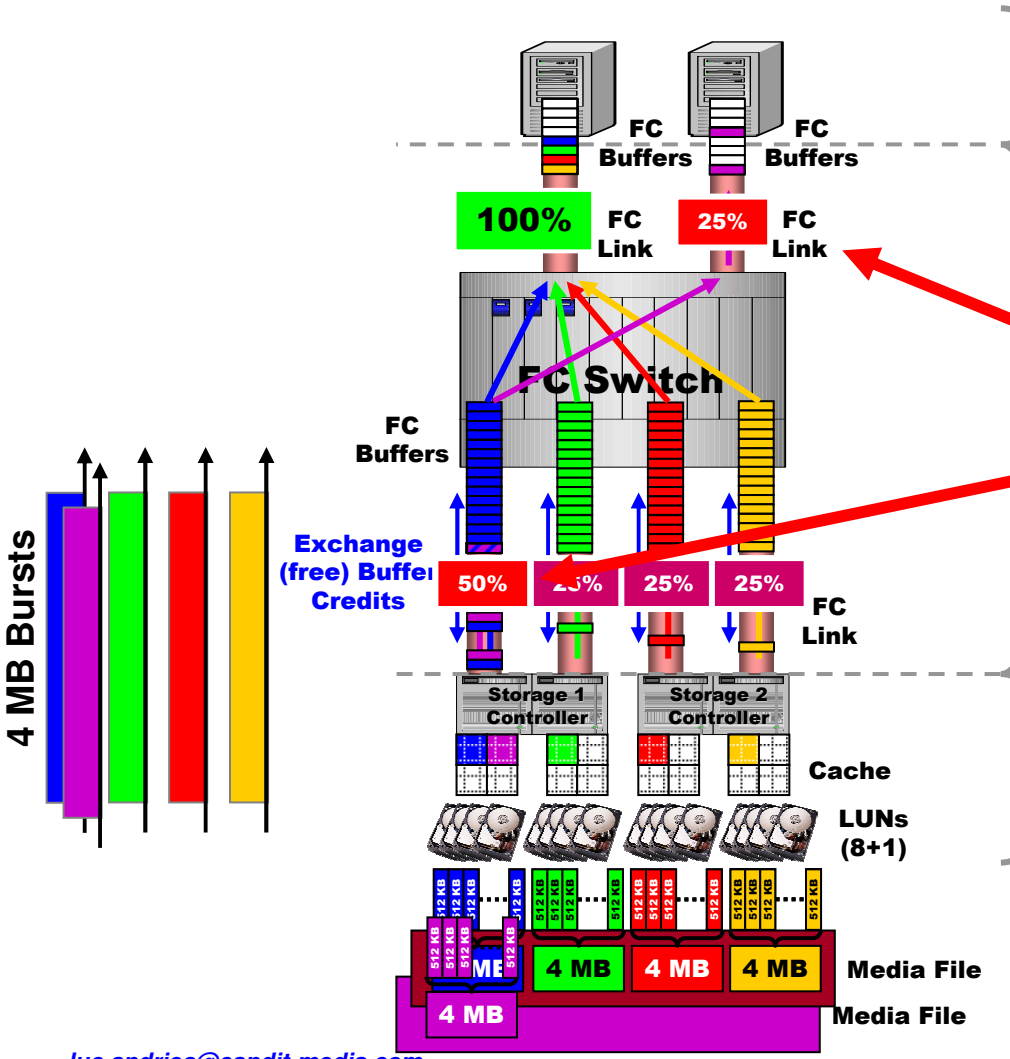


Reading a Media File

- Media File Striping
- Large Block Size/Disk
- RAID 5 (8+1) -> 4MB
- File System READ
- READ from Disk -> Cache
- 4 MB Traffic Bursts
- Buffer_to_Buffer Credits
- Data Transfers (4:1)
- Flow Control Kicks in
- No Packet Loss!!!

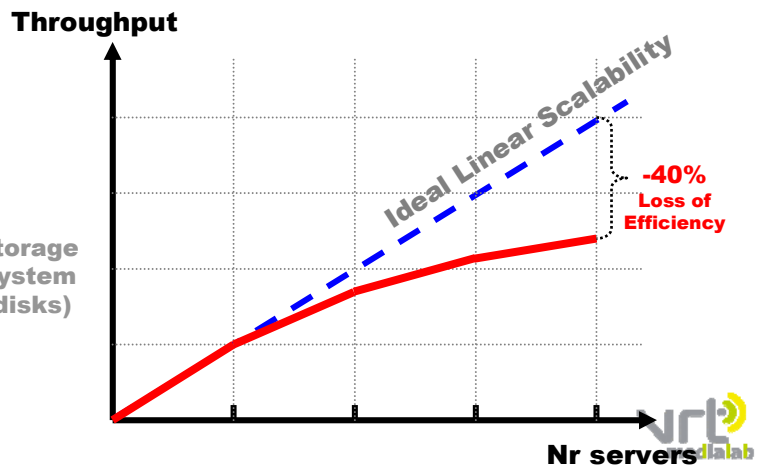
luc.andries@candit-media.com

Media Traffic Interference



Reading Multiple Media Files

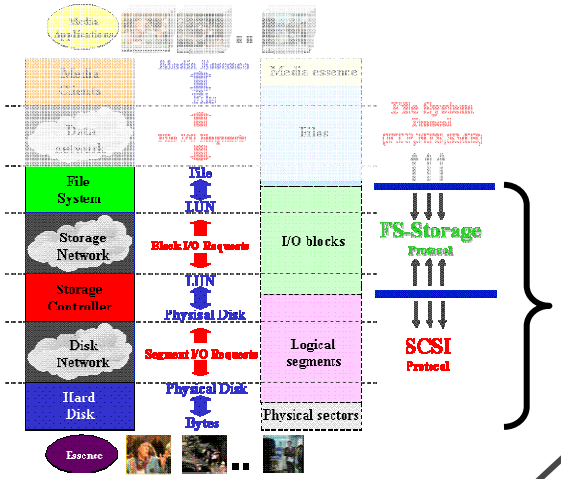
- 2nd Server
- Reading a 2nd File
- Sharing the 'Blue' Link
- 4 MB Traffic Bursts
- Link Level Flow Control Throttles both Flows
- Traffic Interference



FC Does NOT Scale for Media

luc.andries@candit-media.com

Some Conclusions

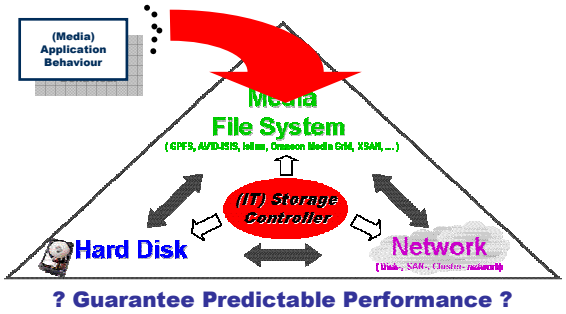


- Access 'Large' Files
- At High Throughput
- In a Multi-User/Multi-Process Context
- At Sustained Load

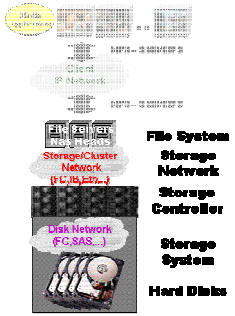
?! Guaranteed Predictable Performance ?!

luc.andries@candit-media.com

(Media) Application Behaviour



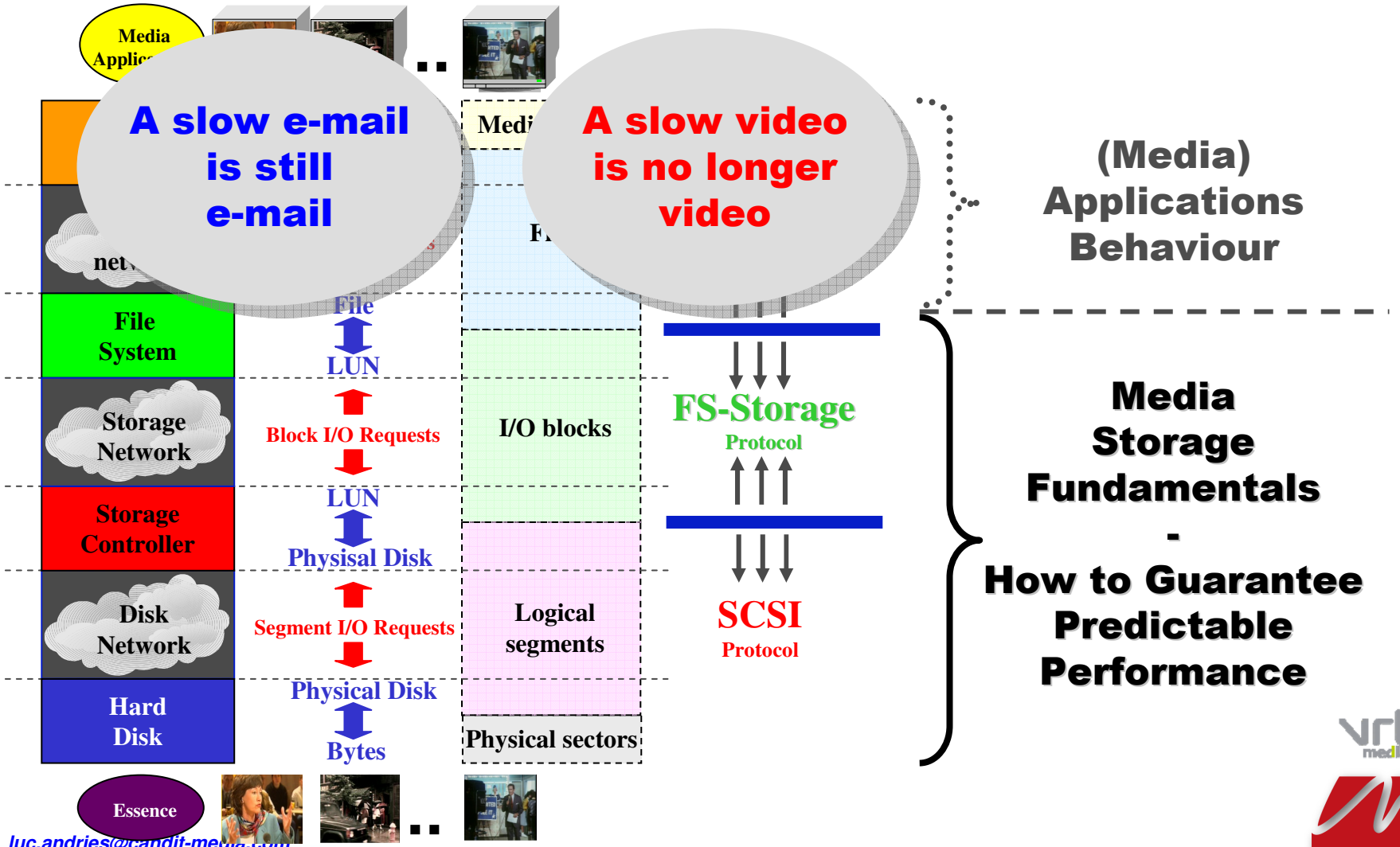
Media Application Behaviour
~



- Depends upon Application
- Depends on File system Protocol used
- Depends on Network Protocol used
- Depends on File System Intelligence
- Depends on the Storage
- Depends on the Network

luc.andries@candit-media.com

Final Conclusion



luc.andries@candit-media.com

CandIT Media

Thank You !

Luc Andries
luc.andries@candit-media.com