SIAV ACADEMY
FONDAZIONE

La Conservazione a Lungo Termine: Tecnologie e Soluzioni Organizzative

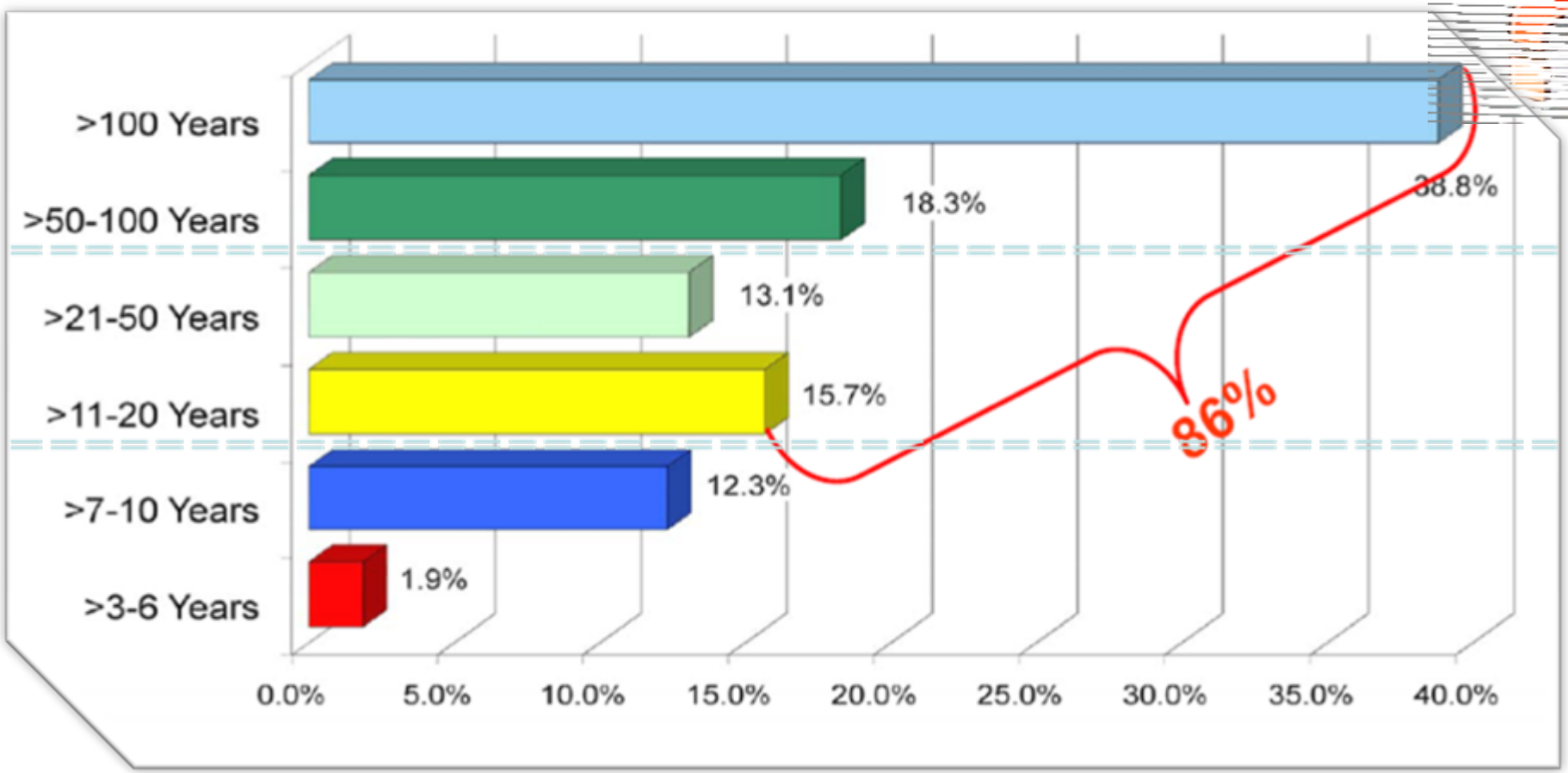
Vincenzo Gambetta

(Roma 8 luglio 2011)

Cosa significa a lungo termine?

[SNIA – 100 Year Archive Requirements Survey]

- *Per quanto tempo dovete conservare i vostri dati?*



Fonte: SNIA – 100 Year Archive Requirements Survey – Gennnaio 2007

Tempi di conservazione obbligatori per la documentazione

	Anni di Conservazione			
	IVA	Imposte Dirette	Esigenze contabili	Esigenze Commerc.
Austria	Da 7 a 22	7	7	7
Belgio	7	5	10	10
Danimarca	Da 5 a 10	5	5	5
Finlandia	6	6	6	N/D
Francia	6	6	6	6
Germania	10	10	10	10
Irlanda	5	5	8	N/D
Italia	4	4	10	10
Olanda	Da 7 a 9	7	N/D	7
Norvegia	10	10	10	10
Portogallo	10	10	10	10
Spagna	4	4	6	6
Svezia	7	7	10	N/D
Svizzera	Da 10 a 20	10	10	10
Regno Unito	6	6	6	N/D

La Conservazione della Memoria Digitale

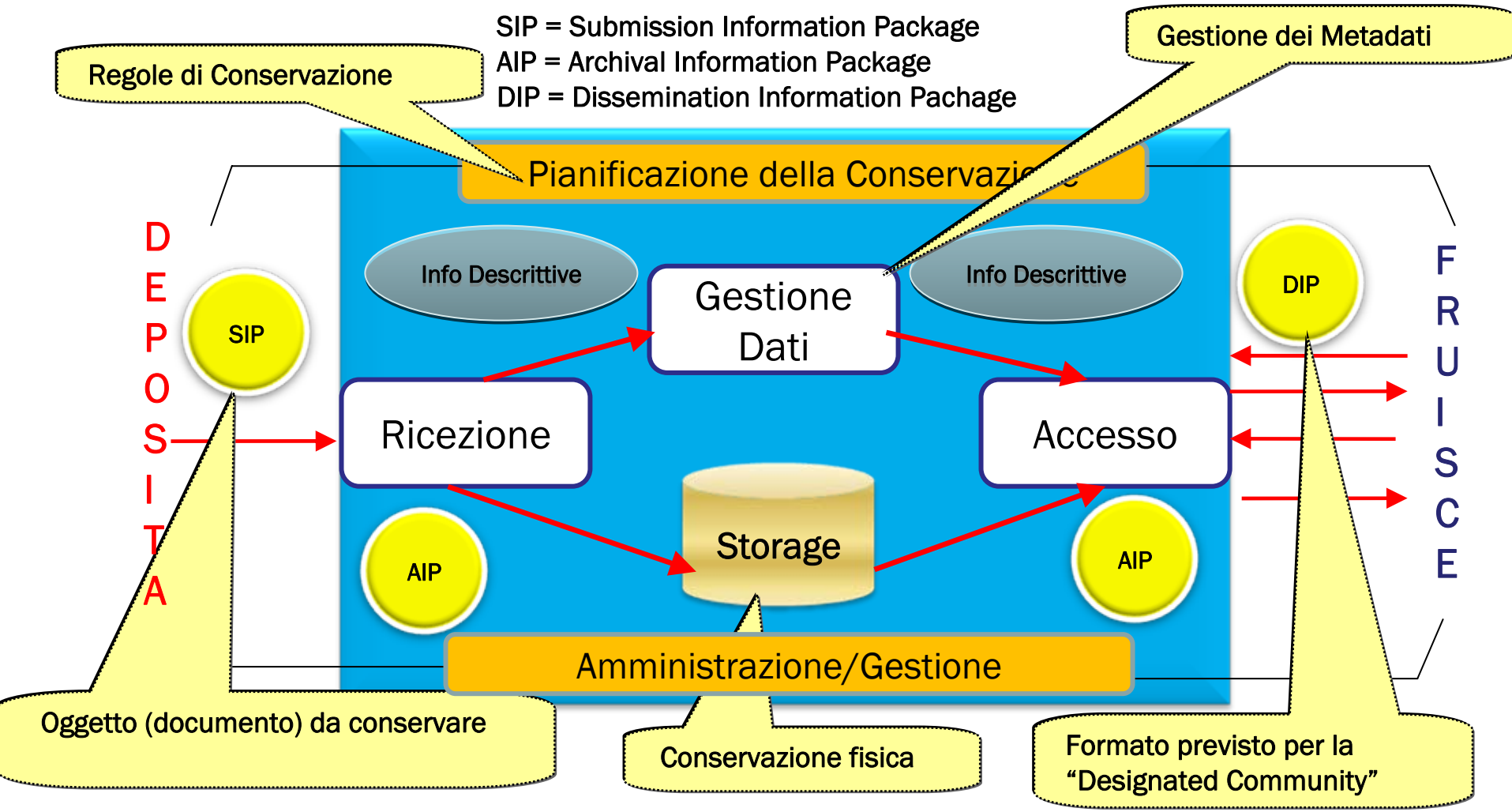
I. Integrità delle registrazioni fisiche

II. Formato

III. Affidabilità e Protezione

IV. Metadati e Regole per gestire l'intero Ciclo di Vita delle Informazioni

Il Modello ISO di "Open Archival Information System"



Fonte: OAIS reference model (ISO:14721:2002).

L'evoluzione della delibera AIPA/CNIPA

Decreto MEF 23/1/2004 - D. L. vo 52 - Recepimento direttiva 2001/115/CE

- AIPA n° 42/2001
  - Conservazione Digitale
  - Supporto di Memorizzazione a Tecnologia Laser
  - Documenti Analogici (Originali e Copie)
  - Documenti Informatici (Firmati Digitalmente)
  - Conservazione per Gruppi di Documenti
  - Riversamento Sostitutivo

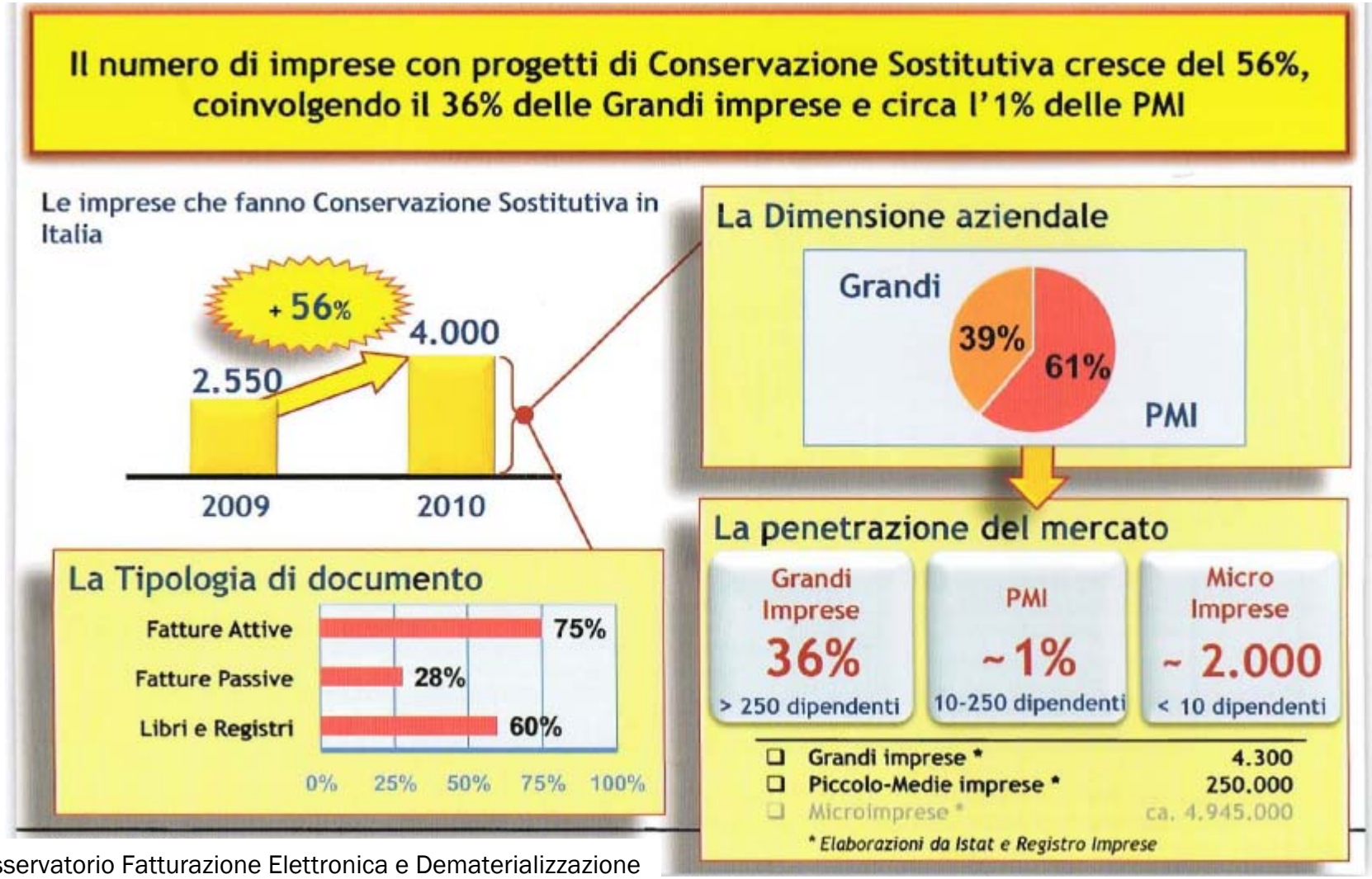
- CNIPA n° 11/2004
  - Conservazione Sostitutiva
  - Definizione di Documento Informatico (*Confusione*)
  - Calcolo dell'impronta dei documenti

Direttiva Europea n° 93/1999 sulle Firme Elettroniche

- AIPA n° 15/1994
  - Archiviazione Ottica Sostitutiva
  - Incompleta
  - Impostata su Standard (inapplicabili, obsoleti)

- AIPA n° 24/1998
  - WORM Fisico
  - Documenti Cartacei (Originali e Copie)
  - Documenti Formati all'origine su supporto informatico
  - Individualità di ogni singolo documento
  - Complessa, ma fattibile

La diffusione della Conservazione Sostitutiva «oggi»



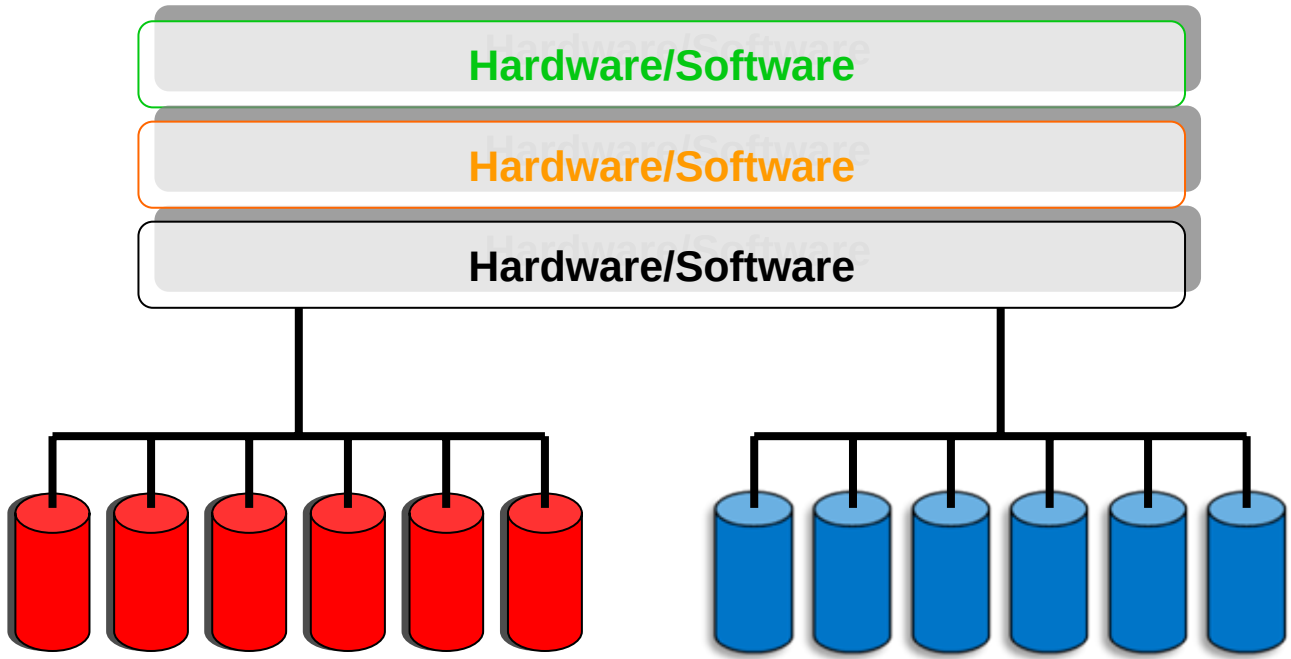
Osservatorio Fatturazione Elettronica e Dematerializzazione – Politecnico di Milano – Maggio 2011

Supporti di memorizzazione “Consumer” e “Professionali”

- I supporti consumer rappresentano una scelta solo apparentemente più economica
  - I CD, DVD, Blu-ray e HD-DVD registrabili, sono stati pensati per applicazioni d’intrattenimento o per distribuire quantità di informazioni a costi ridotti
- L‘adozione di supporti non professionali è limitato e nella maggior parte dei casi coinvolge organizzazioni di piccole dimensioni e con volumi ridotti
  - non si ritiene di dover suggerire modifiche nell’adozione di supporti non professionali per quanto fin ora realizzato.
- Per le realizzazioni future si reputa che l’adozione di supporti non professionali in situazioni con limitati volumi e periodi di conservazione brevi possa essere ancora ritenuta accettabile se accompagnata da opportune misure di prudenza.

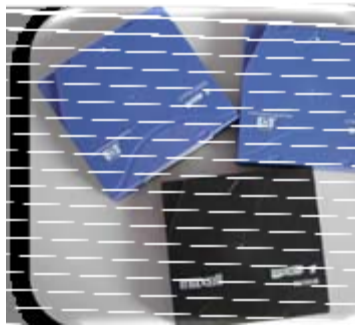
Sistemi Storage

- Sono il risultato dell'aggiunta, alle funzionalità (di base) dei dispositivi di memorizzazione (es.: HDD), di strati HW e SW per ottenere un sistema:
  - ad alta affidabilità, di prestazioni elevate e facile da gestire
- Negli ultimi 50 anni circa si sono modificati per soddisfare l'evolversi delle esigenze degli utenti

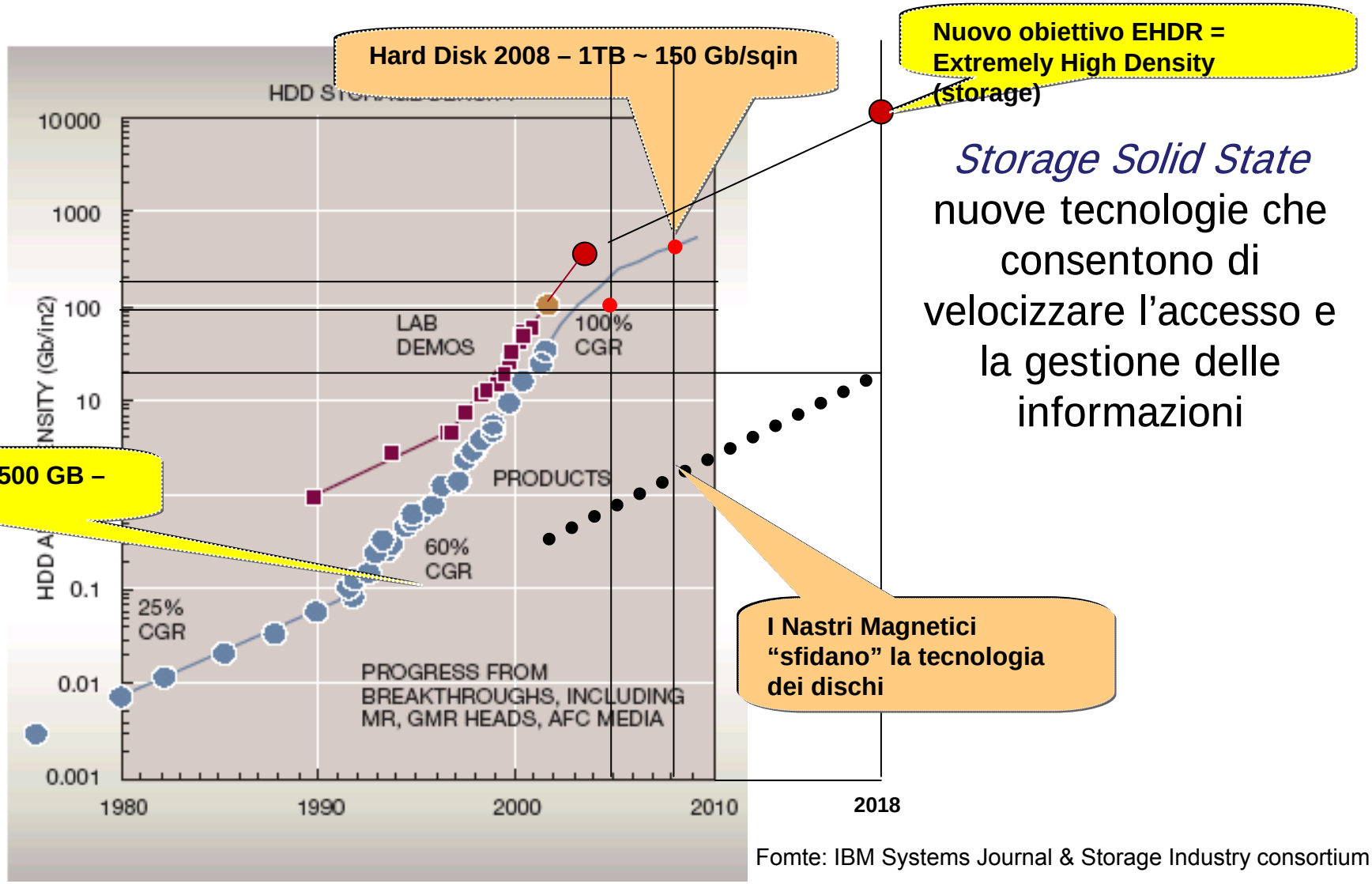


I Supporti Professionali

- Dischi Magnetici e Nastri Magnetici,
  - In genere sistemi hardware complessi e spesso dotati di funzionalità software avanzate
    - Supporti fissi (Dischi magnetici)
    - Supporti rimovibili (Nastri Magnetici)
      - la Cartuccia può mantenere inalterato il suo contenuto per oltre 25 anni
    - Ciclo di vita economico
      - Dischi Magnetici → 3 – 5 anni
      - Nastri → oltre 20 anni per le librerie, 5 anni per le cartucce



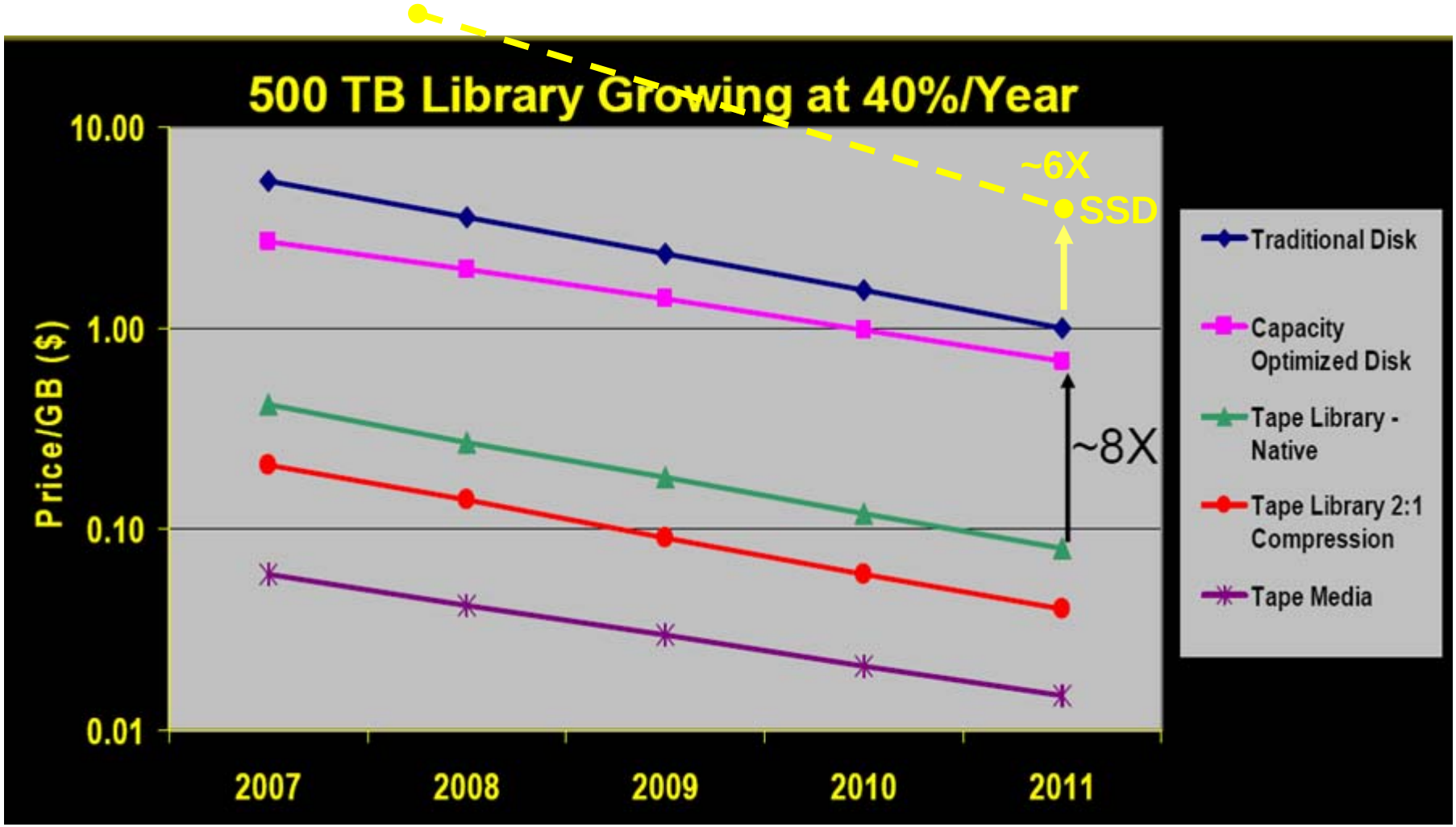
Andamento della densità dei supporti



Storage Solid State
nuove tecnologie che consentono di velocizzare l'accesso e la gestione delle informazioni

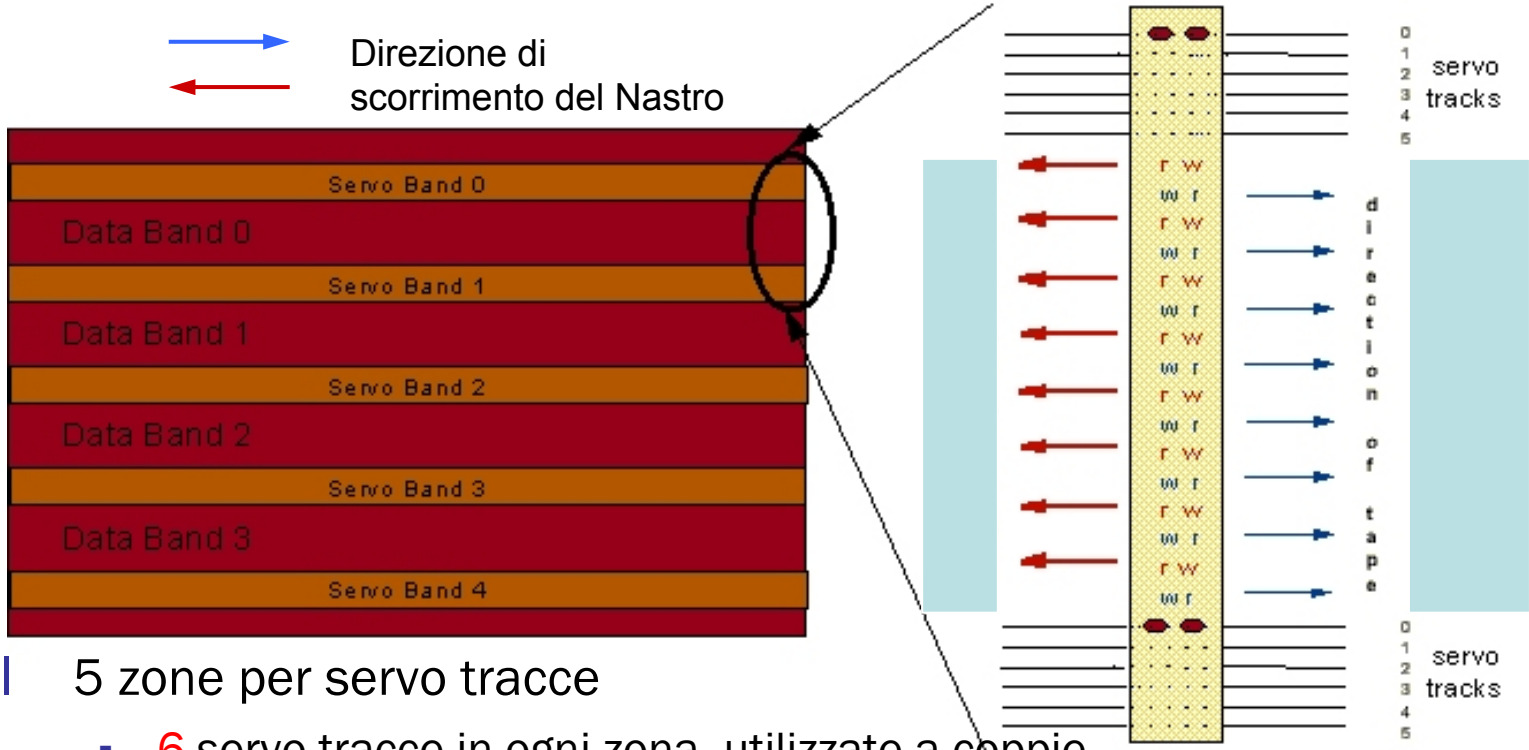
Fomte: IBM Systems Journal & Storage Industry consortium

Andamento dei Costi delle Memorie



Fonte: Information Storage Industry Consortium (INSIC) – 2008

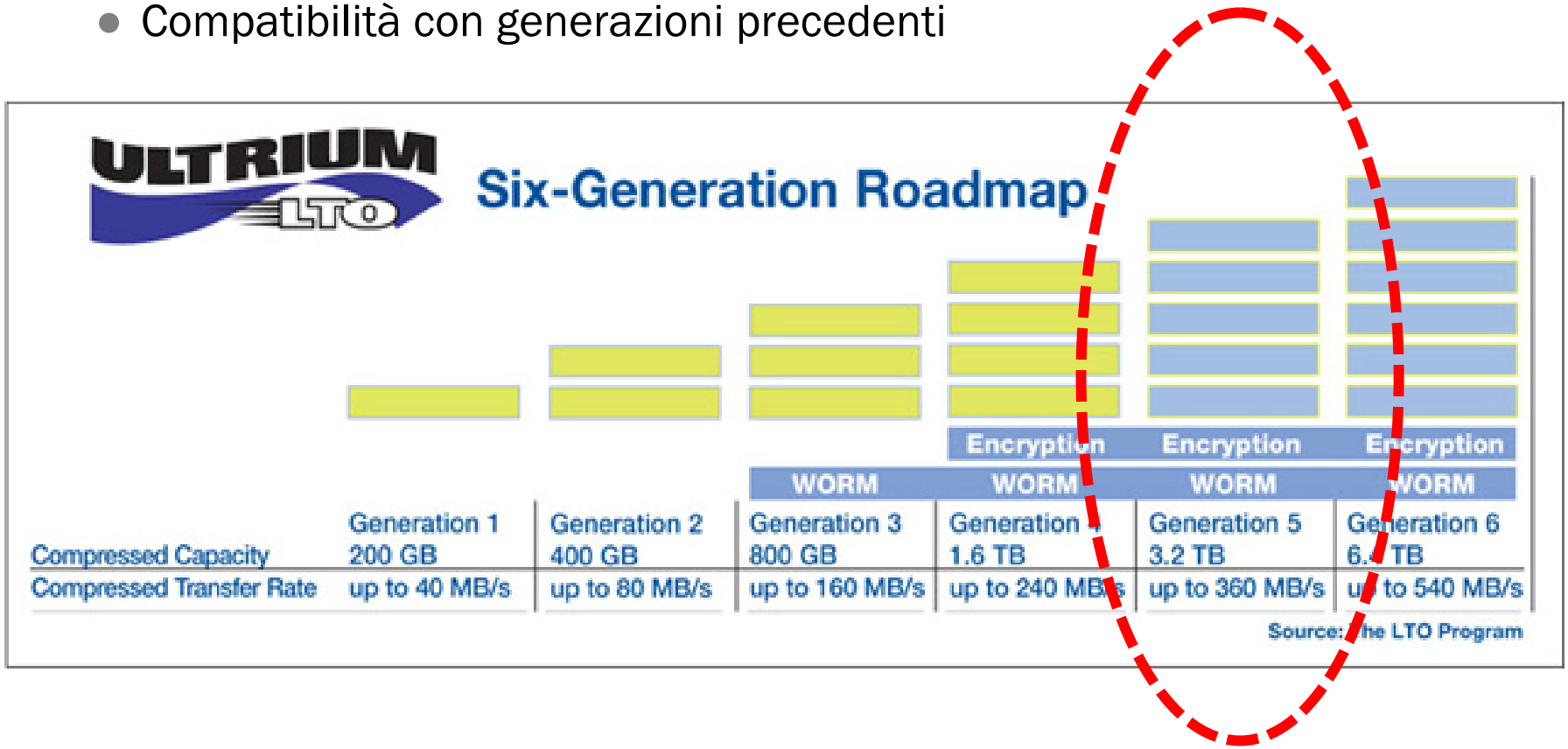
LTO ULTRIUM



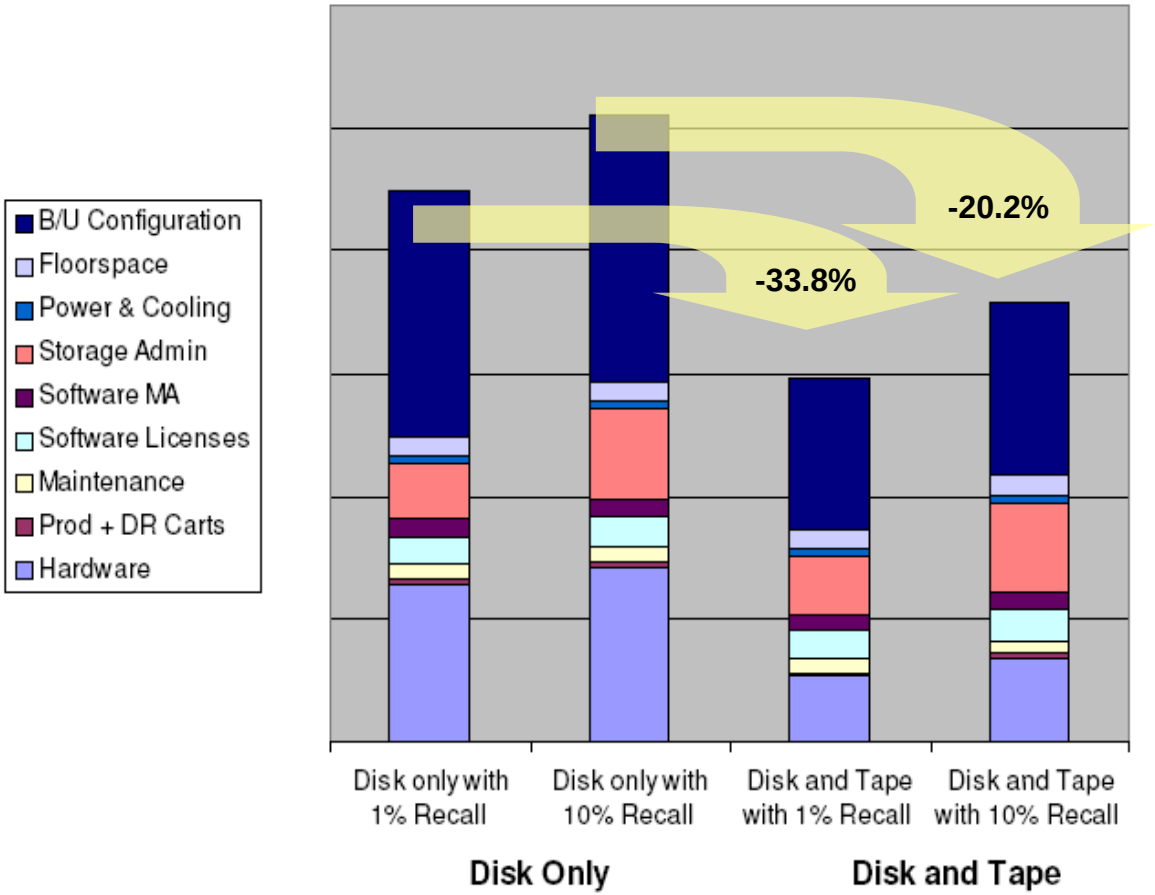
- 5 zone per servo tracce
  - 6 servo tracce in ogni zona, utilizzate a coppie
  - La testina segue le tracce dati con precisione anche in presenza di piccole stirature o piegamenti della banda magnetica
- 4 zone per la memorizzazione dei dati (Per un totale di 384 tracce)
  - 8 tracce alla volta per ogni direzione → Si passa quindi alla seconda coppia di servo tracce

Roadmap Ultrium

- Supportato da più di 30 aziende del settore
  - Nuova generazione ogni 2 anni
  - Compatibilità con generazioni precedenti



TCO (7 anni) di un progetto di Archiviazione



Fonte: Optimizing Data Retention and Archiving – Robert Frances Group - Novembre 2007

Ipotesi

- **Recall rate: la quantità di dati che si ipotizza saranno richiamati dagli archivi** (**1 or 10%**).
- **Data retention period: in questa analisi 7 anni**.
- **High performance access duration: quanto a lungo i dati staranno dui dischi ad alte prestazioni** (assunto = **6 mesi**).
- **Contingency capacity: capacità in eccesso per far fronte a esigenze non previste** (assunto = **10%**).
- **Utilizzo dei Dischi e delle Cartucce Nastro: Assunto 70% and 85% rispettivamente**.
- **Initial load requirements: spazio totale necessario inizialmente per memorizzare i dati** (assunto = **5TB**).
- **Aggiornamento della tecnologia: assunto 3 anni per i dischi and 6 anni per i nastri**.
- **Scelta delle Tecnologie : Nastri LTO4 e dischi SATA 750GB**
- **Storage administration: si è assunto che una persona possa gestire fino a 5TB**

Archiviazione di lungo periodo: TCO ed evoluzione tecnologica

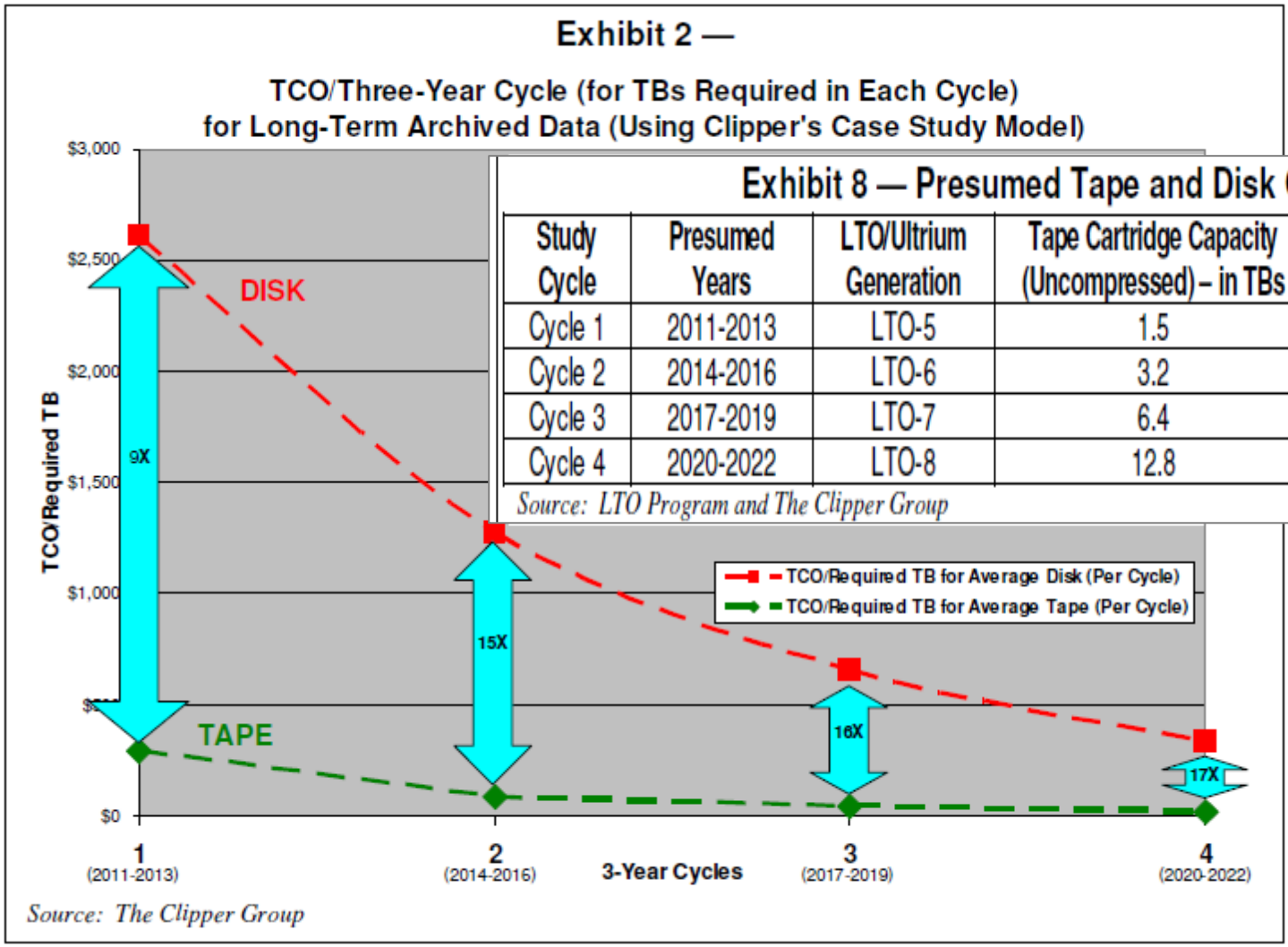


Exhibit 8 — Presumed Tape and Disk Capacities

Study Cycle	Presumed Years	LTO/Ultrium Generation	Tape Cartridge Capacity (Uncompressed) – in TBs	SATA Drive Capacity (Uncompressed) – in TBs
Cycle 1	2011-2013	LTO-5	1.5	2
Cycle 2	2014-2016	LTO-6	3.2	4
Cycle 3	2017-2019	LTO-7	6.4	8
Cycle 4	2020-2022	LTO-8	12.8	16

Source: LTO Program and The Clipper Group

Tecniche e strategie per conservare nel tempo le Informazioni Digitali

- Le informazioni in formato digitale e si affidano per la loro "sopravvivenza" ai dispositivi fisici
- Protezione contro la perdita delle registrazioni
  - nel periodo in cui le informazioni sono "attuali"
  - nel periodo di conservazione
- Uso appropriato delle Tecnologie Raid
  - Compromessi tra prestazioni e affidabilità
  - Nel lungo periodo il tipo di RAID da adottare sarà uno che premia la protezione contro la perdita delle informazioni (elevata ridondanza)
- Verifica automatica di quanto registrato
- Duplicabilità delle informazioni digitali
  - Opportune strategie di Replica (le stesse informazioni sono mantenute aggiornate in siti differenti) o Copia
  - Back-up, Recovery/Restart
  - Disaster Recovery

Conservare = Copiare, copiare, copiare!

Raid (Redundant Array of Independent Disks)

- Individua modalità di registrazione di dati in più punti (registrazioni ridondanti) di un sistema con più dischi fisici (supporti, testine, controller) ed è visto dal Sistema Operativo come un unico disco logico
- Il risultato di registrare dati su più dischi può avere come risultato di consente il voluto equilibrio tra
  - Prestazioni in lettura o i lettura
  - Livello di protezione/l'affidabilità del sistema
  - Spazio utilizzabile
  - Velocita di ricostruzione dei dati
- Dal punto di vista dell'affidabilità si possono considerare due situazioni:
  - Mirroring – ogni registrazione è duplicata su due dischi fisici differenti. Costosa, ma molto efficace [la rottura simultanea dei due dischi è molto improbabile]
  - Controllo di parità – registra il risultato di un algoritmo (Checksum) che consente di ri-calcolare una porzione di informazione se questa viene persa
- Livelli di RAID

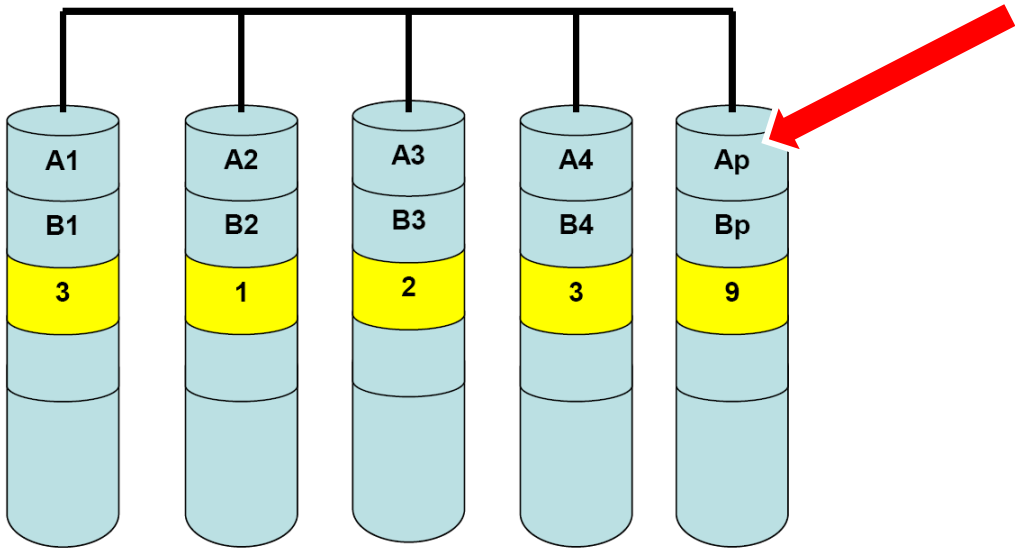
RAID

- RAID-0 – [→ RAID-10] Consente il massimo livello di prestazioni in quanto le registrazioni non avvengono in sequenza, ma sono eseguite in parallelo nel numero massimo di partizioni consentite. In tale situazione si ha il minimo livello di affidabilità possibile [minino 2 dischi]

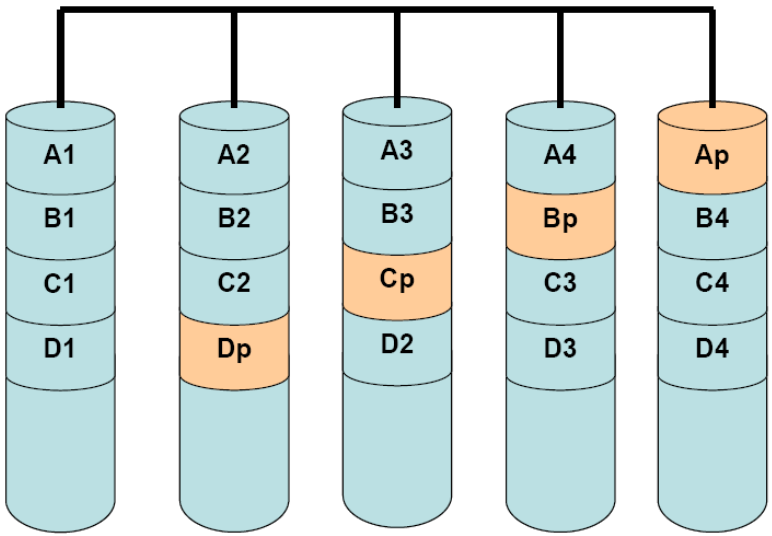
- RAID-1– Si tratta di Mirroring. I dati sono duplicati su almeno due dischi fisici. Le prestazioni in scrittura non variano da quelle di un solo disco, mentre quelle in lettura possono trarre vantaggio dalla lettura simultanea del numero di copie eseguite. L'affidabilità è massima

RAID

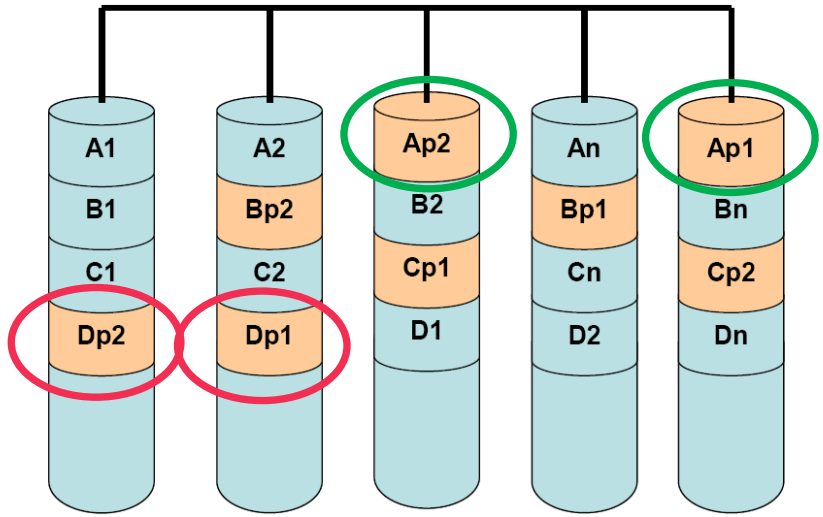
Raid 4



Raid 5



Raid 6



Limiti ed evoluzioni del RAID

- Quando in un sistema RAID (ad esempio 5) un disco viene meno il sistema lo rigenera in un altro disco, coinvolgendo tutti gli altri.
- Questo incremento di attività aumenta la probabilità di altri guasti di un ordine di grandezza → aumenta il rischio di perdita di dati
- 20 anni fa la ricostruzione dei dati si misurava in minuti, ma con le capacità odierne (TB) le ricostruzioni possono richiedere giorni … … … se non settimane.
- Varie le tendenze in atto per superare il problema …

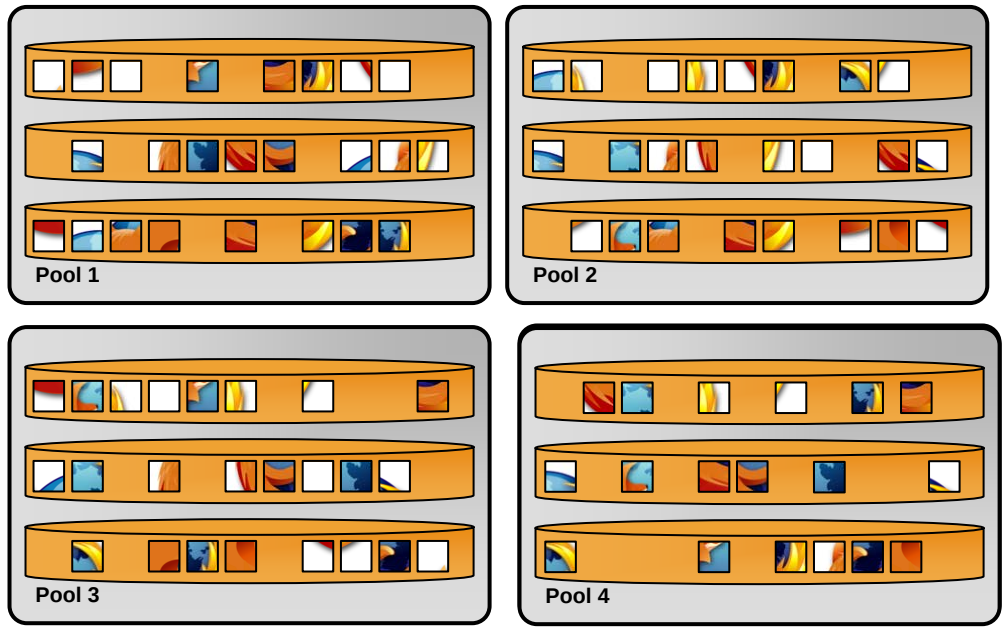
Evoluzioni del RAID

- Triple mirror
- Innovazione
  - Declustering → variante di raid 10 che impiega algoritmi intelligenti per distribuire causalmente i blocchi (mirror) sull'intero array (può ricostruire un RAID di 2TB in 30 minuti).
  - Network Raid che adotta i nodi storage al posto di drive disco: ciò consente di distribuire i blocchi di un volume logico; ciò consente a livello di singolo cluster di avere anche 4 mirror.
    - Controlli automatici di controllo delle registrazioni consentono di copiare e riparare i dati in un altro nodo prima che il guasto si verifichi.

Declustered RAID

Risolve il problema dell'affidabilità e delle prestazioni (Input/Output Per Secondo)

- Gruppi di dischi a basso costo (JBOD[1])
- Suddivisione Stocastica delle stripe
- Scalabilità
- Maggiore affidabilità

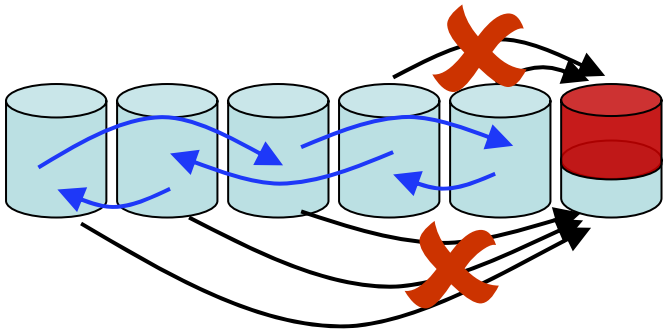


[1]Concatenazione di dischi per combinare un insieme più dischi fisici in un grande disco virtuale al fine di far sembrare l'insieme come un singolo disco.

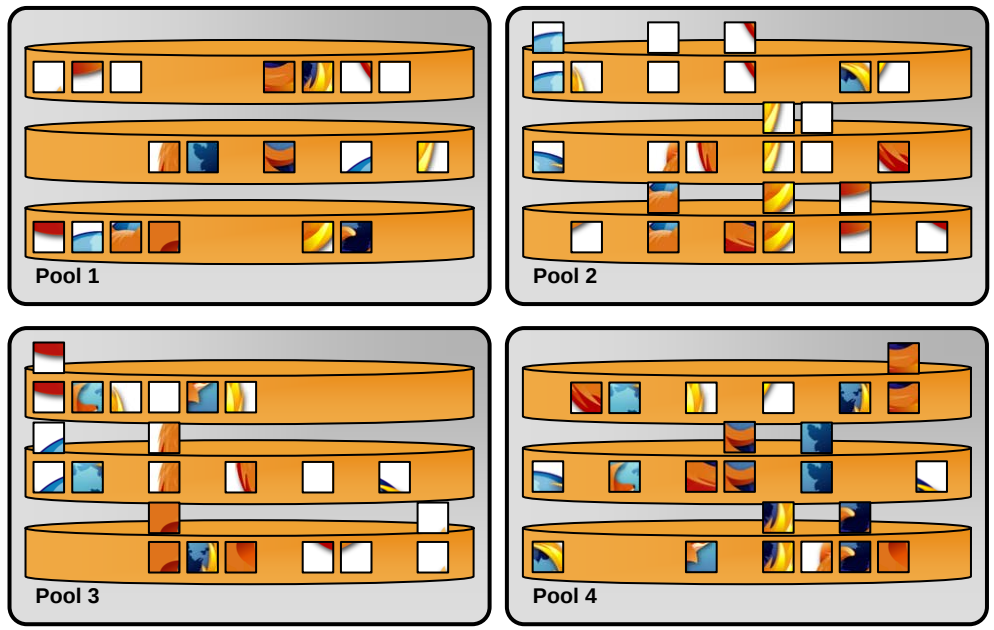
Declustered RAID

Risolve il problema dell'affidabilità e delle prestazioni (Input Output Per Secondo)

- Non necessita di dischi "Spare" e non necessita di ricostruzione



[Gruppo di dischi che viene meno]



Evoluzioni del RAID

- Trasformazione
  - Sistemi dotati di un'auto diagnosi continua e riparazione automatica che assomigliano molto a quelle adottate in fabbrica (analisi dei cicli dell'alimentazione, ricalibratura dei componenti, ...)
    - se il risultato è negativo la ricostruzione avviene solo a livello dei singoli settori che hanno dato errore.

Evoluzioni del RAID

- Cambio di paradigma→ Approccio completamente diverso dal RAID.
  - Erasure Codes - Suddivide i dati in “pezzi” (chunk) non auto-consistenti, con informazioni per ricostruire l’intero insieme di dati a partire da un subset di pezzi {< è il numero di pezzi necessari, > saranno le info addizionali].
    - I chunk possono essere distribuiti in differenti posizioni all’interno di un data center, città, zona geografica o sparsi per il modo.
    - È insensibile ai disastri naturali o ai guasti tecnologici in quanto per ricostruire le info sono sufficienti sono una porzione di pezzi. Può anche supportare guasti multipli simultanei.
    - Particolarmente adatto per Cloud o storage distribuito perché non ha bisogno di replicare i dati e può distribuire i dati su più località geografiche

Virtualizzazione

- Creazione di una “versione virtuale” di una risorsa fisica (es. divisione di un disco fisso in partizioni logiche)
- Un ulteriore impiego, di contro, è il riunire più unità storage in rete in un insieme che venga visto come un’unica entità.
- Meccanismi più avanzati di virtualizzazione permettono la ridefinizione dinamica tanto delle caratteristiche della risorsa virtuale, quanto della sua mappatura sulle risorse reali.
- Tra i vantaggi è la possibilità di adottare qualunque tecnologia di storage, tra le più eterogenee - per tipologia e fornitore senza preoccuparsi (*eccessivamente*) di farle convivere e di cosa accadrà nel momento in cui tale tecnologia diventerà obsoleta e dovrà quindi essere rimpiazzata.
- Virtualizzare significa, in pratica, separare o dividere quello che l’utente o l’applicazione “vede” ed utilizza per accedere ai propri dati da quello che è la realtà hardware/software che eroga il servizio.

Tiered Storage vs "Hierarchical Storage Management" (HSM)

- Differenze Sostanziali
- Hierarchical Storage Management ottimizza lo spazio utilizzato (prestazioni di accesso ai dati e frammentazione dei file, dati consultati raramente).
- I dati sono spostati su dispositivi più economici e con prestazioni inferiori, mentre i dati più richiesti sono memorizzati su dispositivi più rapidi.
- L'obiettivo principale del Tiered Storage è quello di ridurre i costi globali dello storage, ottimizzando i tempi d'accesso ai dati rilevanti per il business e ottenendo la necessaria conformità alle norme di riferimento.

Cloud

- Insieme delle tecnologie informatiche che consentono l'utilizzo remoto di risorse hardware (storage e risorse di calcolo) o software distribuite
  - *virtualizzazione delle risorse, resa possibile da tecnologie, strumenti, potenza elaborativa nonché dalla capacità e sicurezza delle reti oggi disponibili*.
- Il termine è abusato ed utilizzato in contesti e con significati differenti.

Cloud - Classificazione

- In funzione del modello
  - Privato, Pubblico o Ibrido.

- In funzione del tipo di risorsa utilizzata
  - Software as a Service: l'utente utilizza delle applicazioni
  - Platform as a Service: l'utente può sviluppare e distribuire su un'infrastruttura Cloud applicazioni
  - Infrastructure as a Service: L'utente può acquisire ed utilizzare capacità elaborativa, storage, network ed altre risorse di base.

Cloud

- Principali Timori/Problemi
  - Sicurezza
  - Disponibilità
  - Prestazioni
  - Timore che il modello di pagamento risulti troppo oneroso
  - Carenza di interoperabilità
  - Timori nei confronti della Personalizzazione
  - Interoperabilità (?)
    - Integrazione con applicazioni in house
    - Eventuale rientro

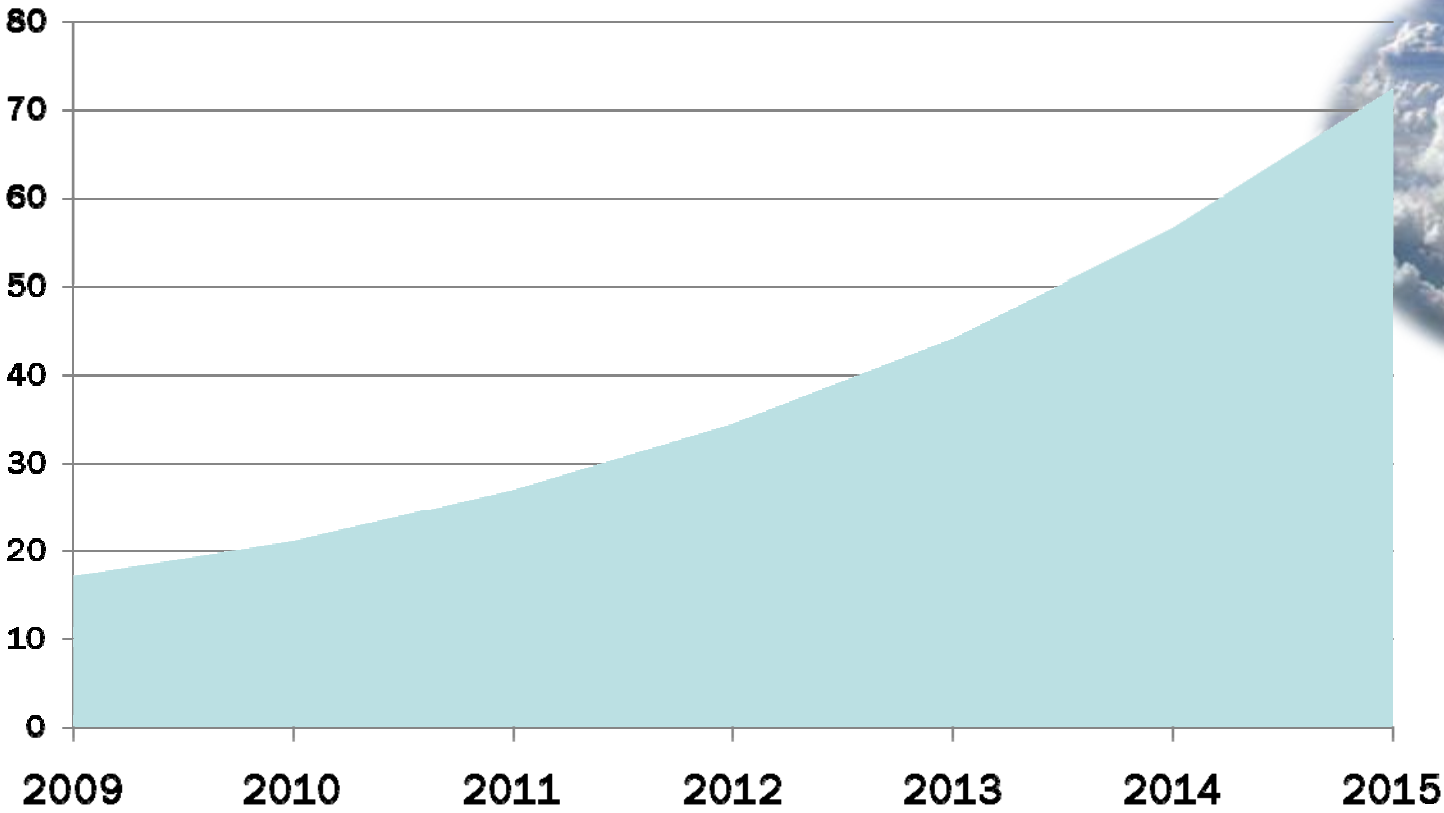
Cloud - Impatto

- Se l'adozione delle tecnologie Cloud proseguirà come prevedibile, l'impatto economico dell'impiego del Cloud Computing sui cinque principali paesi europei è valutato, entro il 2015, sui 177,3 miliardi di Euro, così suddivisi: Francia 37,4, Germania 49,6, Italia 35,1, Spagna 25,2 e Regno Unito 30,0.
- Vantaggi economici indiretti dell'adozione del cloud originati, nei cinque paesi:
  - incremento del PIL per un valore complessivo di 280 miliardi di Euro,
  - l'occupazione diretta ed indiretta che si genererà potrebbe raggiungere i 2.396.000 di unità.

Fonte Centre for Economics and Business Research (2011)]

Cloud - Impatto

Mercato mondiale del Cloud Computing (in Midi di $)



Tecnologie Sinergiche: Dispositivi Mobili, Reti wireless, Social networkink e Business Analitycs

Fonte IDC (2010/2011)]

Evoluzione tecnologica

- Le tecnologie di memoria digitale sono soggette a continui miglioramenti
- Necessità di trasferire i dati su nuove tecnologie

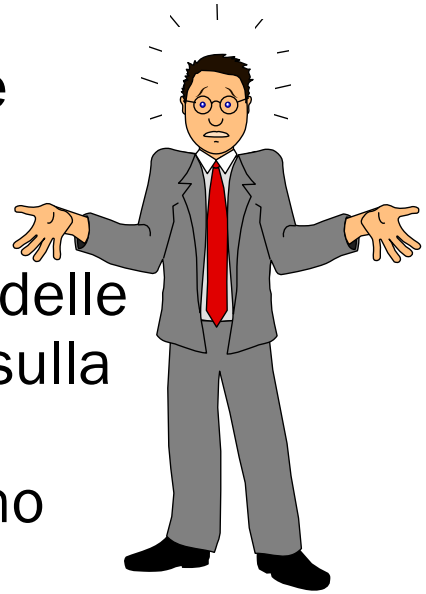
Il processo di riversamento periodico deve

- ***diventare parte delle procedure di gestione del sistema,***
- ***essere ben documentato,***
- ***essere operativo un sistema di verifica del corretto completamento di tale operazione***

  - L'adozione di tecnologie di memoria standardizzate facilita il riversamento dei dati.

- I sistemi intelligenti che disaccoppiano la tecnologia dalla memorizzazione dei dati sono sistemi di virtualizzazione che consentono di utilizzare memorie differenti e di renderli visibili al server e/o alle applicazioni come un'unica risorsa, indipendentemente dal modo in cui sono utilizzati fisicamente

Compiti del Responsabile della Conservazione



- Definire, realizzare, documentare ed aggiornare periodicamente (comunque ad ogni variazione) le procedure di archiviazione, di protezione dati e di protezione da eventi disastrosi
- Verificare periodicamente la corretta esecuzione delle procedure di archiviazione (controlli a campione sulla validità dei dati registrati)
- Verificare periodicamente mediante prove (almeno annualmente) la capacità di ripristino dei dati coinvolgendo il personale addetto
  - Documentare l'esito delle prove e delle eventuali azioni per la correzione e/o miglioramento del processo
- Verificare periodicamente (almeno ogni 3 anni) l'obsolescenza tecnologica dei dispositivi utilizzati, del formato e del software di accesso ai dati al fine di tempificare opportunamente le attività di riversamento.

I Formati

- “Formato” - modo in cui l’informazione digitale è strutturata e registrata sul supporto di memorizzazione
  - corpo
  - testata o header Formati preferenziali
- Formati compatibili
- Portabilità
- Formati proprietari e formati standard

I Formati

- Formati → in aumento (programmi/piattaforme)
- ~ 1.500 formati (estensioni); per un terzo non è disponibile la descrizione
  - Testi/documenti - (DOC, DOCX, XML, HTML, PDF,...)
  - Immagini - (GIF, JPG, BMP, TIF, CGM, EPS, SVG, ...)
  - Suoni - (MP3, WAV, ...)
  - Filmati - (MPG, MPEG, AVI, WMV,...)
  - Dati binari, più o meno strutturati, come, ad esempio, gli output di strumentazione di laboratorio
  - Eseguibili - (EXE, ...)
  - Calcolo - (XLS, MDB, ...)
  - Archiviazione e Compressione - (ZIP, RAR, ...).
  - E-mail

I Formati e l'operatività

- Consentono e facilitano l'operatività quotidiana
- Vanno valutati in funzione de
  - La diffusione, ossia il numero di persone ed organizzazioni che li adotta
  - La portabilità (meglio se essa è indotta dall'aderenza a standard documentati e accessibili)
  - Le funzionalità che l'utente ha a disposizione per elaborare l'informazione e collegarla ad altre (ad esempio gestione di link)
  - la capacità di occupare il minor spazio possibile in fase di memorizzazione
  - la possibilità di gestire il maggior numero possibile di metadati
  - La capacità dei programmi di gestire contemporaneamente un numero congruo di formati (in funzione delle esigenze dell'utente)
  - La diffusione di visualizzatori che consentono la fruibilità delle informazioni, indipendentemente dalla possibilità di rielaborarle.

I Formati e la Conservazione nel Lungo Periodo

- Caratteristiche che un formato deve avere [per essere "riconosciuto" e utilizzato correttamente nel tempo]
  - **Immodificabilità** – il formato non deve consentire di apportare modifiche al contenuto informativo (a qualunque livello, compresa struttura e modalità di visualizzazione).
  - **Mantenimento della fedeltà di presentazione del contenuto** – ciò non significa semplicemente conservare nel tempo la struttura del documento, ma anche i cosiddetti "font" e tutte le altre caratteristiche utili alla completa comprensione del documento.
  - **Non contenere riferimenti ad altre informazioni o documenti** (ad esempio link ipertestuali) non sotto il controllo della stessa "unità di informazione" o documento.
  - Il formato di conservazione **deve corrispondere fedelmente ad uno standard** (formale o de facto) **interamente descritto e le cui specifiche siano state rese di dominio pubblico senza limitazioni**

- Il processo di conversione dal formato di un documento "operativo ad uno di conservazione deve essere semplice.

I Formati e la Conservazione nel Lungo Periodo

- Alla data sono pochi i formati che soddisfano i requisiti appena illustrati:
  - PDF/A-1, descritto nello standard ISO 19005-1
    - può disporre di uno “strato” XML (in grado di trasportare dati strutturati e metadati)
    - ha la possibilità di “inglobare” nativamente un’eventuale Firma Digitale
  - Formato immagine (Raster) TIFF in versione non compressa o compressa senza perdita di informazione.
    - di questo formato vi sono parecchie versioni, alcune delle quali proprietarie
    - in genere le specifiche sono pubbliche e non soggette ad alcuna forma di limitazione.
    - sono formati molto utilizzati per la conversione in digitale di documenti cartacei.
    - Il loro impiego, se attuano una forma di compressione, va valutato attentamente.
  - JPEG 2000 (ISO/IEC 15444-1)
    - può utilizzare la compressione senza perdita di informazione
    - consente di associare metadati ad un’immagine.
    - diffusione limitata

Altri Formati

- Open XML
  - standard ISO/IEC 29500 - libero
  - elaborazione di testi, presentazioni e fogli di calcolo
  - larga adozione da parte delle Amministrazioni internazionali
  - *embedding* dei font
  - gestione dei metadati
  - componenti del formato sono racchiuse in un pacchetto ZIP
- TXT
  - Specifica della Codifica del Carattere (Character Encoding)
- Documenti a rilevanza sanitaria
  - DICOM (immagini da strumenti diagnostici) ... non è solo un formato ma definisce anche protocolli e altro
  - HL7 - in particolare CDA2 (Clinical Document Architecture)

Firme ELETTRONICHE

p7m (ingloba l'oggetto firmato) e p7s (disgiunta dall'oggetto):

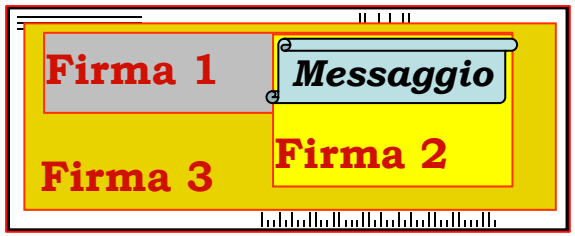
- Singola
- Multipla parallela
- Multipla a controfirma

PDF e XML:

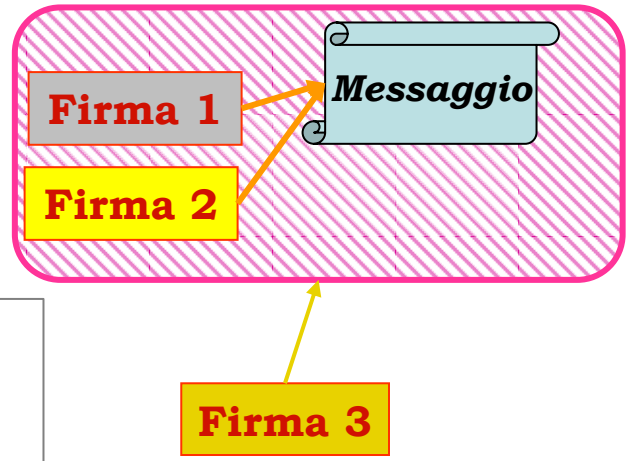
- Singola
- Multipla parallela
- Multipla a controfirma
- Parallela o controfirma con riferimento solo a una parte del documento

Formati di Firma

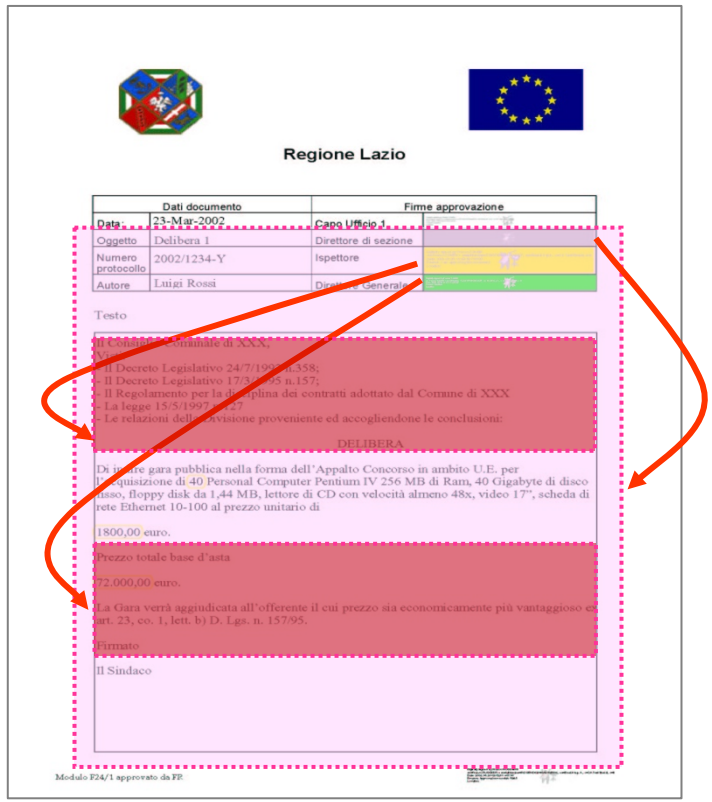
p7m =



p7s =



pdf =



SGM Studio Gambetta S.a.s.
di Gambetta A. & C.
Consulenze & Progetti ICT

Dr. Vincenzo Gambetta

via Salvator Rosa 10, 20851 Lissone (MB)
tel/fax 039.482182 - mobile 335.8035063
v.gambetta@mclink.it

P. IVA e C.F. 02757560962