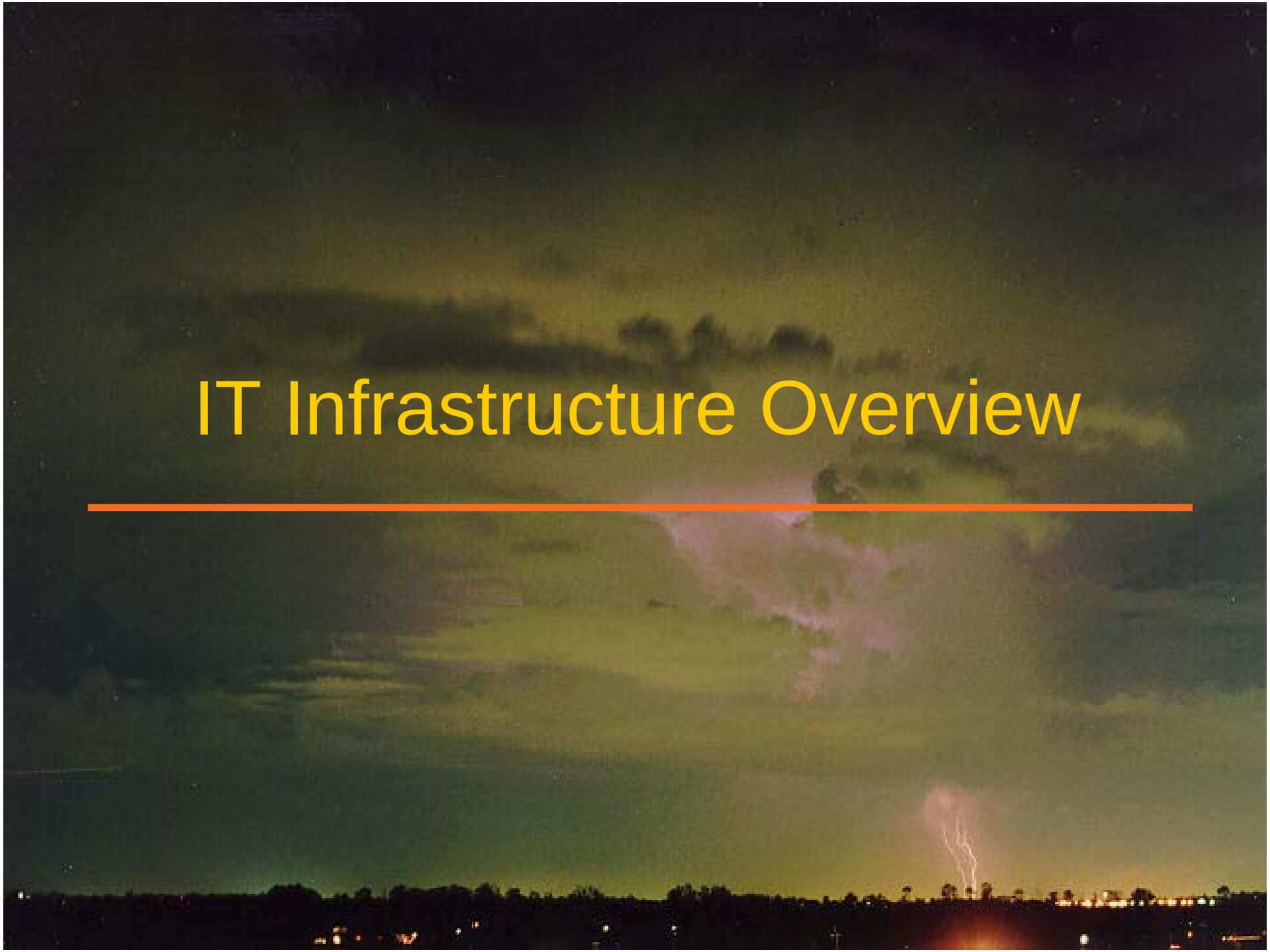
MSC HPC Infrastructure Update

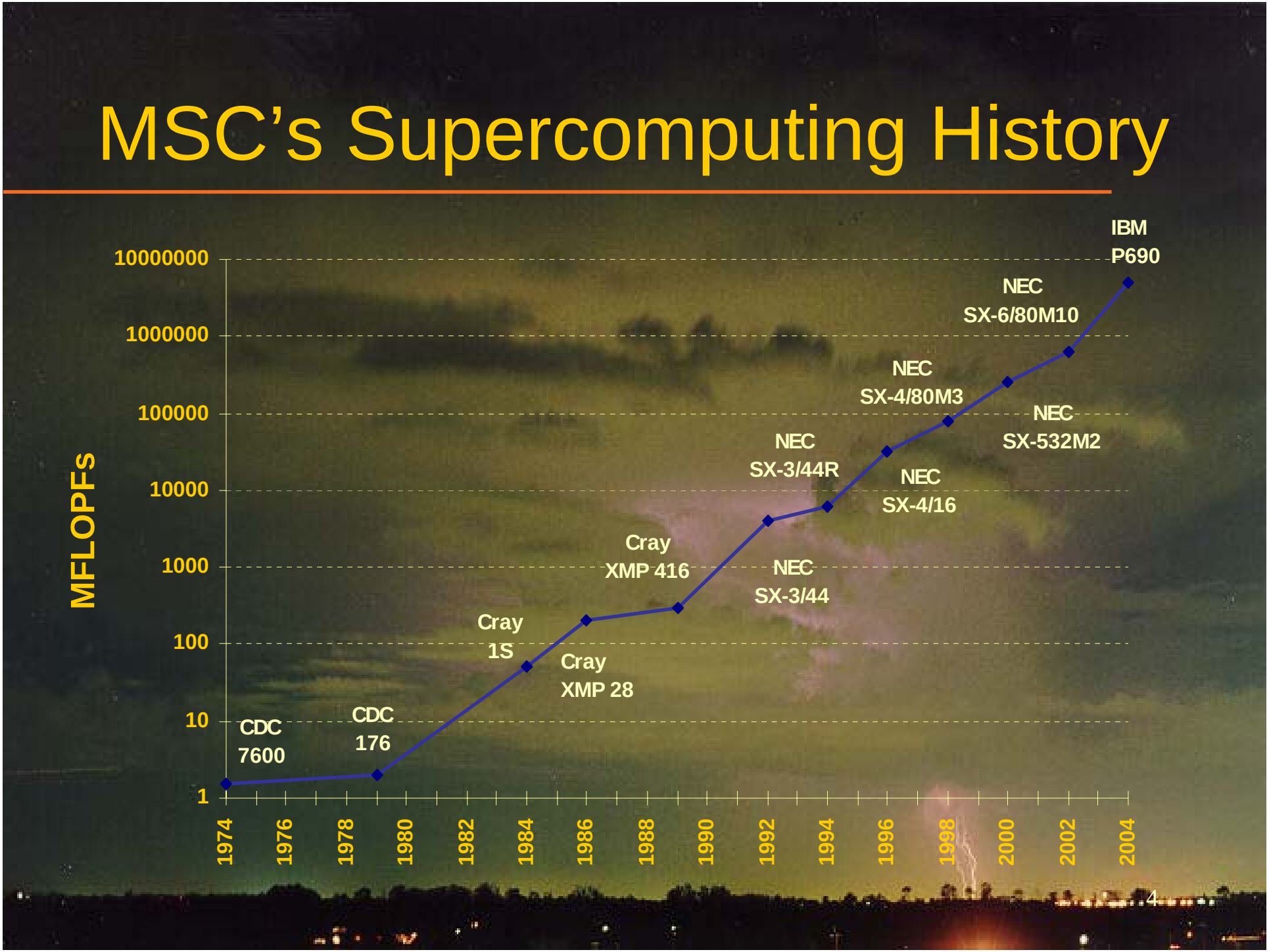
Alain St-Denis
Canadian Meteorological Centre
Meteorological Service of Canada

Outline

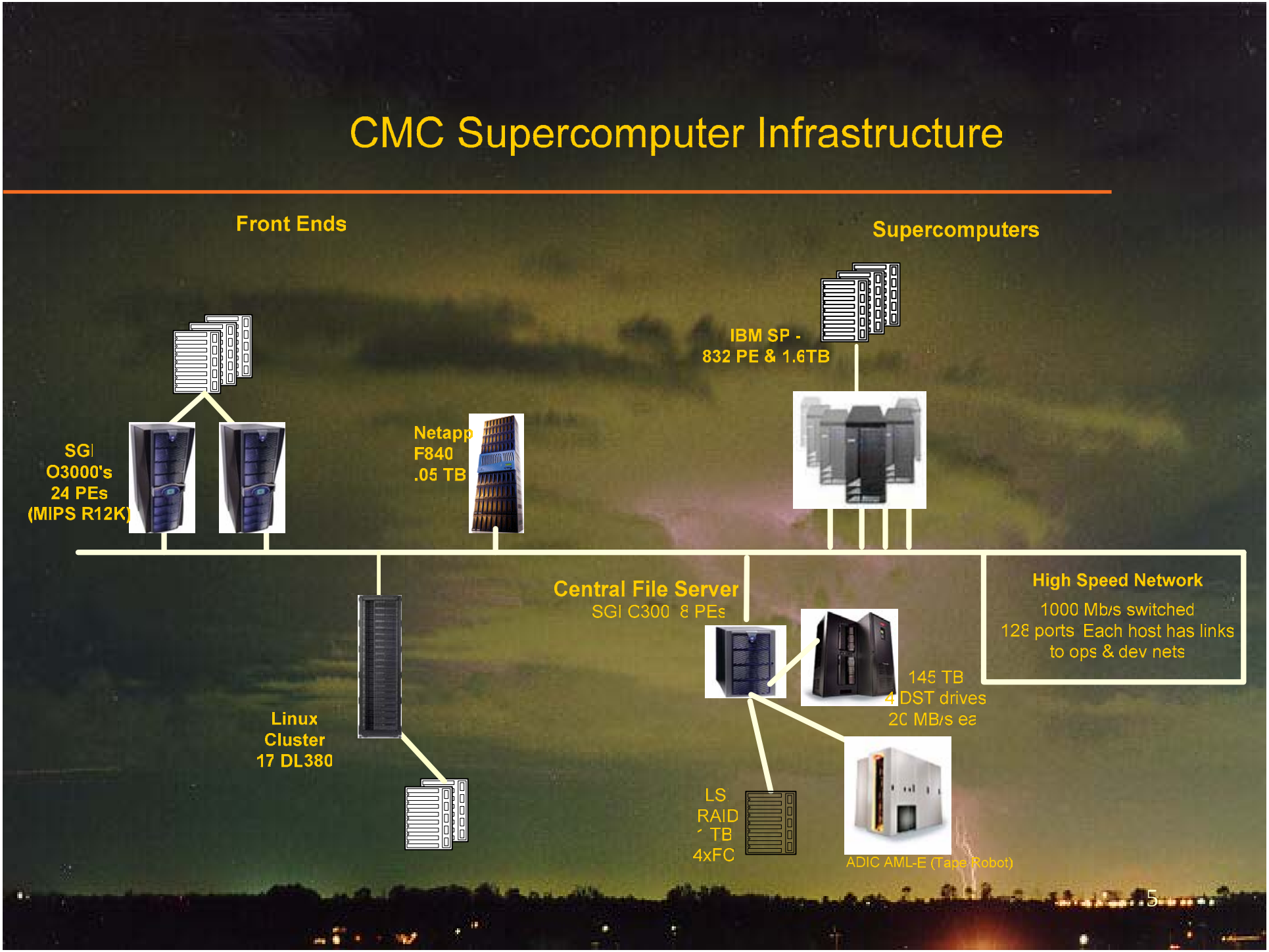
- HPC Infrastructure Overview
- Supercomputer Configuration
- Scientific Direction


IT Infrastructure Overview

MSC's Supercomputing History



CMC Supercomputer Infrastructure

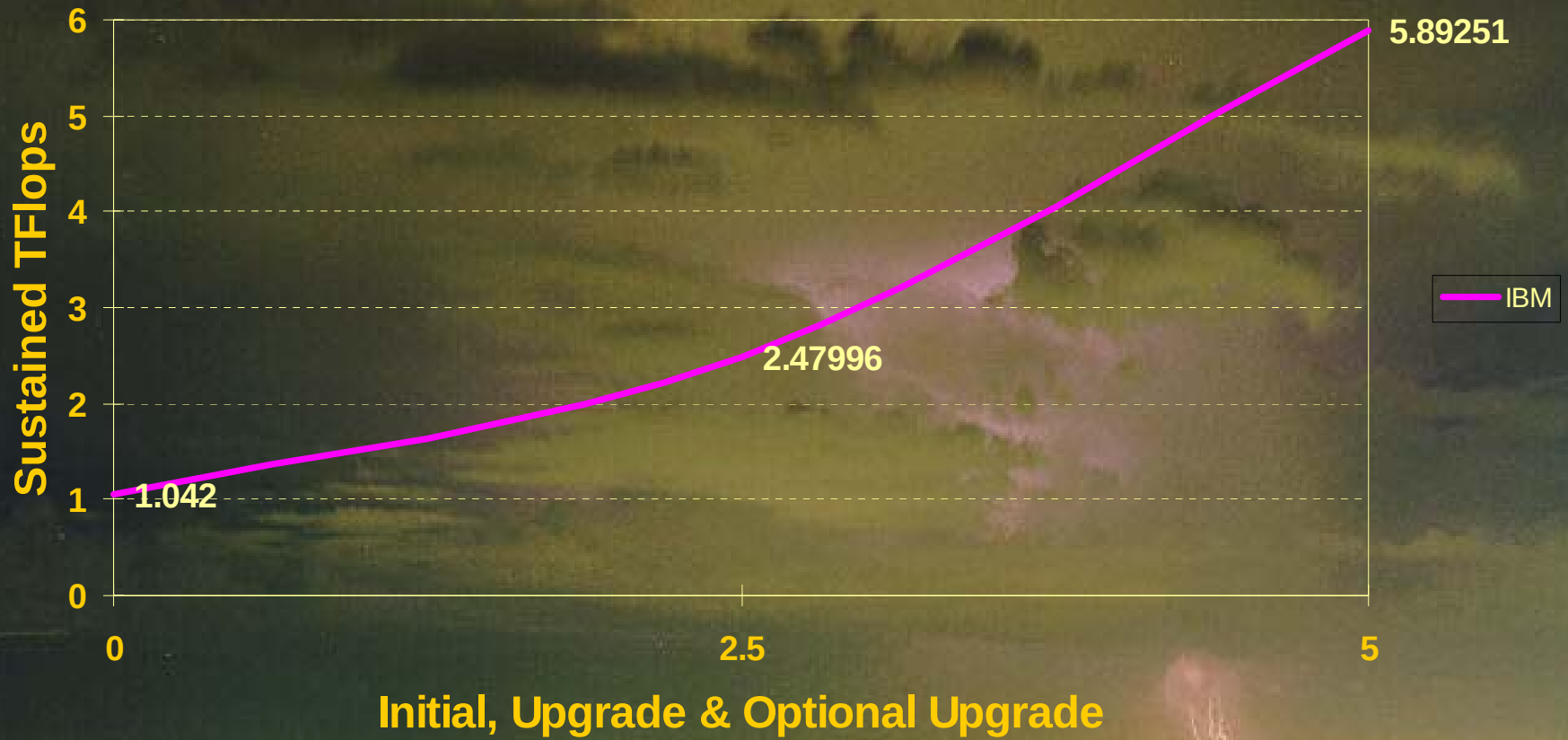


Supercomputer: IBM

IBM

- 128 nodes
- 944 PE
- 2.19 TB Memory
- 15 TB Disks
- 4.825 Tflops (peak)

IBM Committed Sustained Performance in MSC TFlops



Front-End Servers

Twin Servers: Production & R&D systems

- Production Server & R&D:
- SGI 3000: 20 R14000 600MHz CPUs, 20GB RAM
- Together both systems have ~13TB disk
- Production Server: 80% Batch

Front-Ends: Linux Cluster

- Configuration: 17 nodes with a 3TB (raw) SAN
- Each node:
  - 2 processors, P3 - 1.26GHz / P4 – 2.4GHz
  - 2 GB memory
  - 1 FC HBA
  - 1 GE
- Compute nodes batch-only
- Linux Clusters are growing like weed…

Automated Archiving System

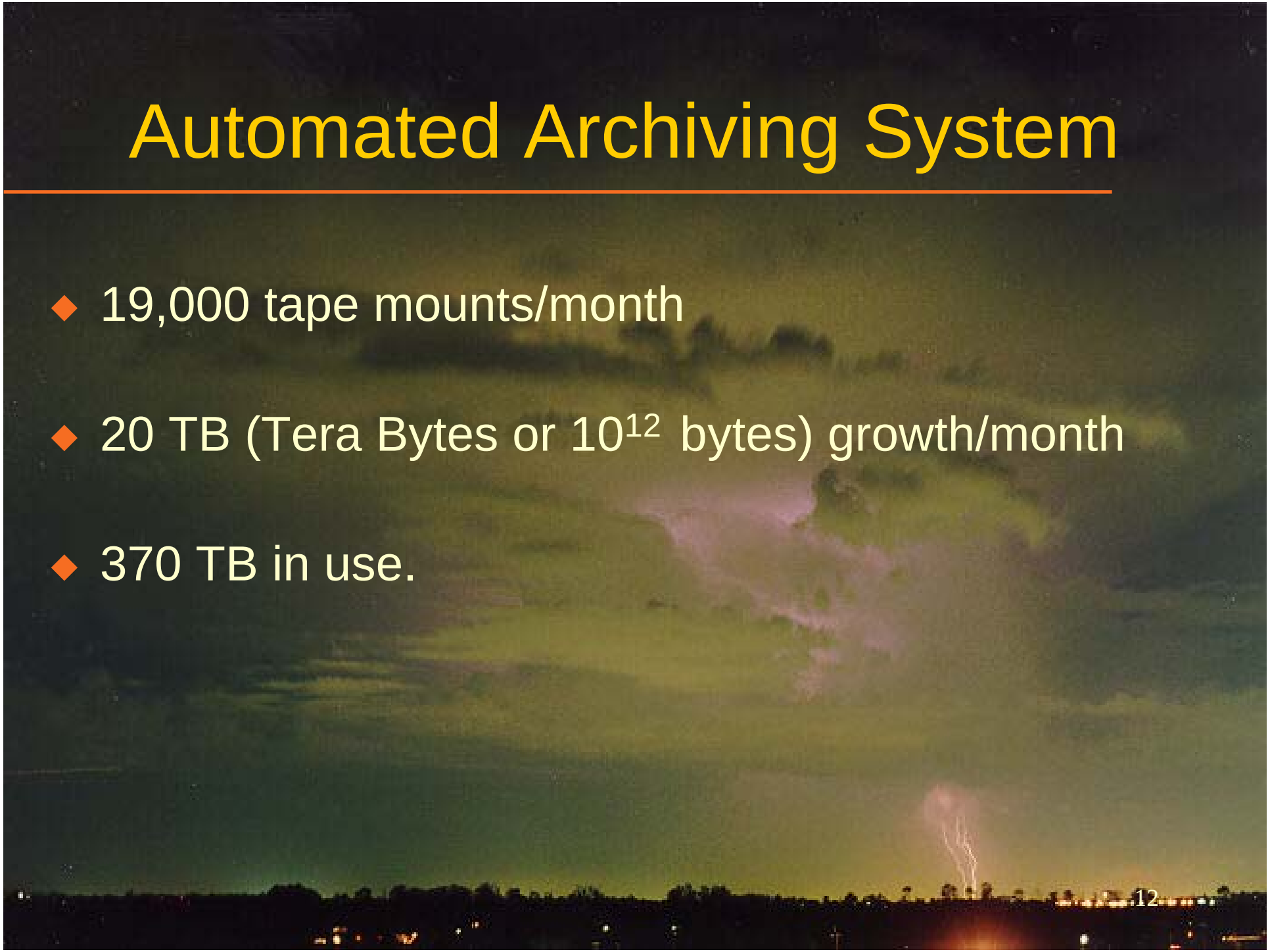
- SGI Origin 300 with 8 processors & 8 GB memory
- 1 TB of FC disk drives
- 144 TB Grau AML/E robot with 4 AMPEX DST-312
- 70 TB Scalar 10K with 10 LTO tape drives
- HSM is FileServ (ADIC) + home grown backup management software

Automated Archiving System

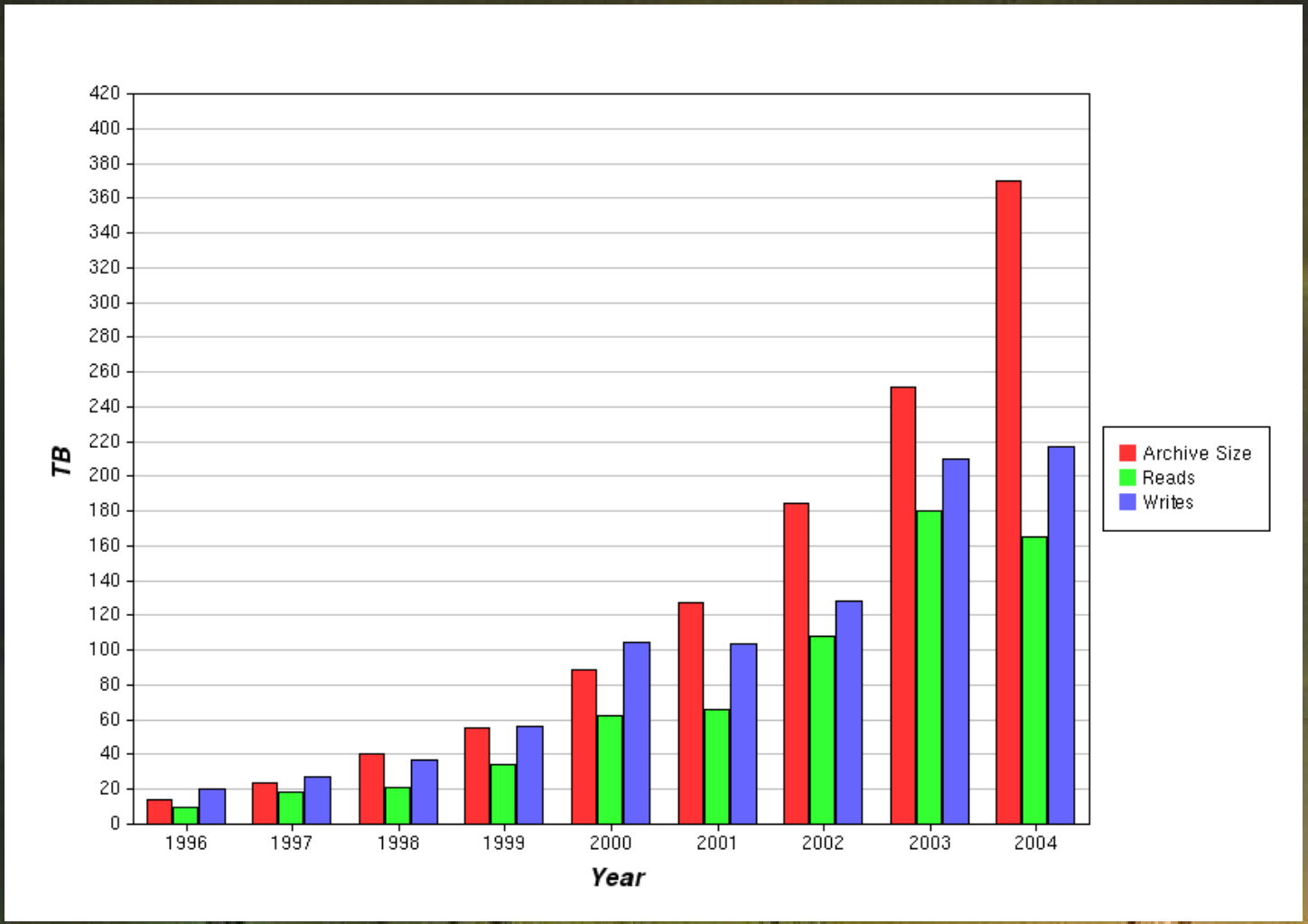
- Main NWP archive (operational run data, R&D)
- Main climate research archive
- Main backup for all systems including the climate archive.
- Limited satellite and radar imagery archive

Automated Archiving System

- 19,000 tape mounts/month
- 20 TB (Tera Bytes or 10^{12} bytes) growth/month
- 370 TB in use.



Automated Archiving System




Supercomputer
Configuration

Software Levels

- AIX 5.1 ML6
- PSSP 3.5
- LoadLeveler 3.1
- GPFS 2.1

Software Maintenance

- Two hours window for maintenance: we use alternate disk installs.
- Try to stay up to date, but not obvious.
- Still not completely familiar with IBM's software distribution methods…

GPFS Configuration

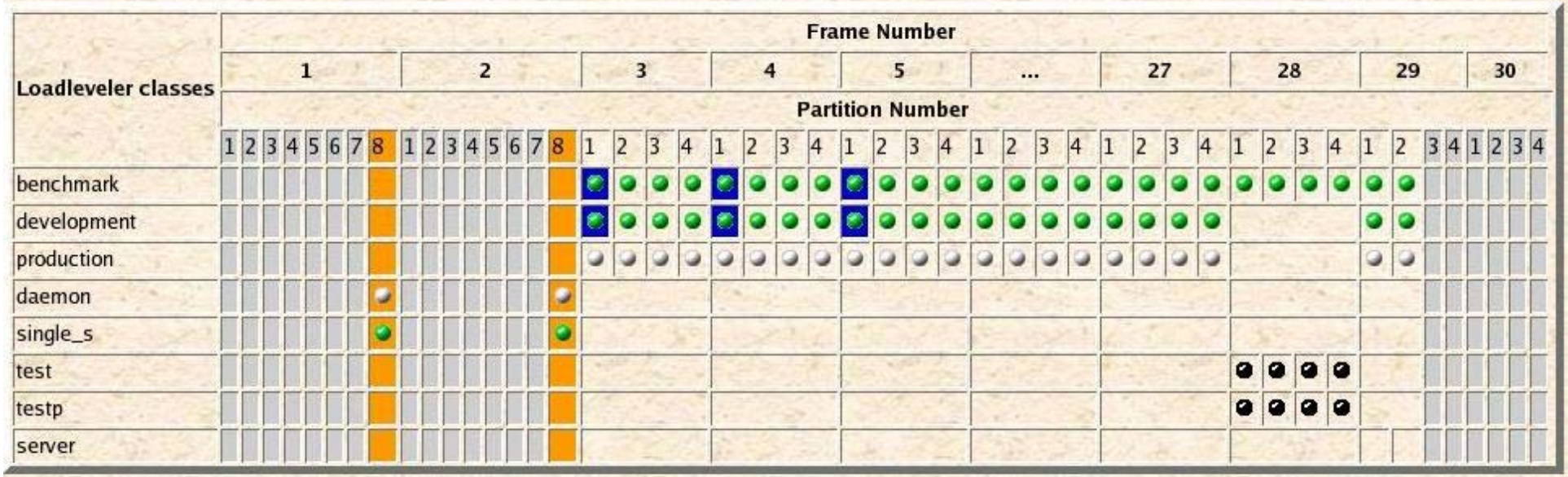
- 18 VSD Servers, dual FC, 17 FASTt700 controllers.
- 4 VSD reserved for production i/o (1 file system).
- 14 remaining servers for development (15 file systems).

Batch Sub-system

- Use of gang scheduler only for preemption.
- Two main classes: production, development.
- Negotiator and Scheduler running on the service nodes.
- One schedd.

Batch Sub-system

LoadLeveler Configuration Information

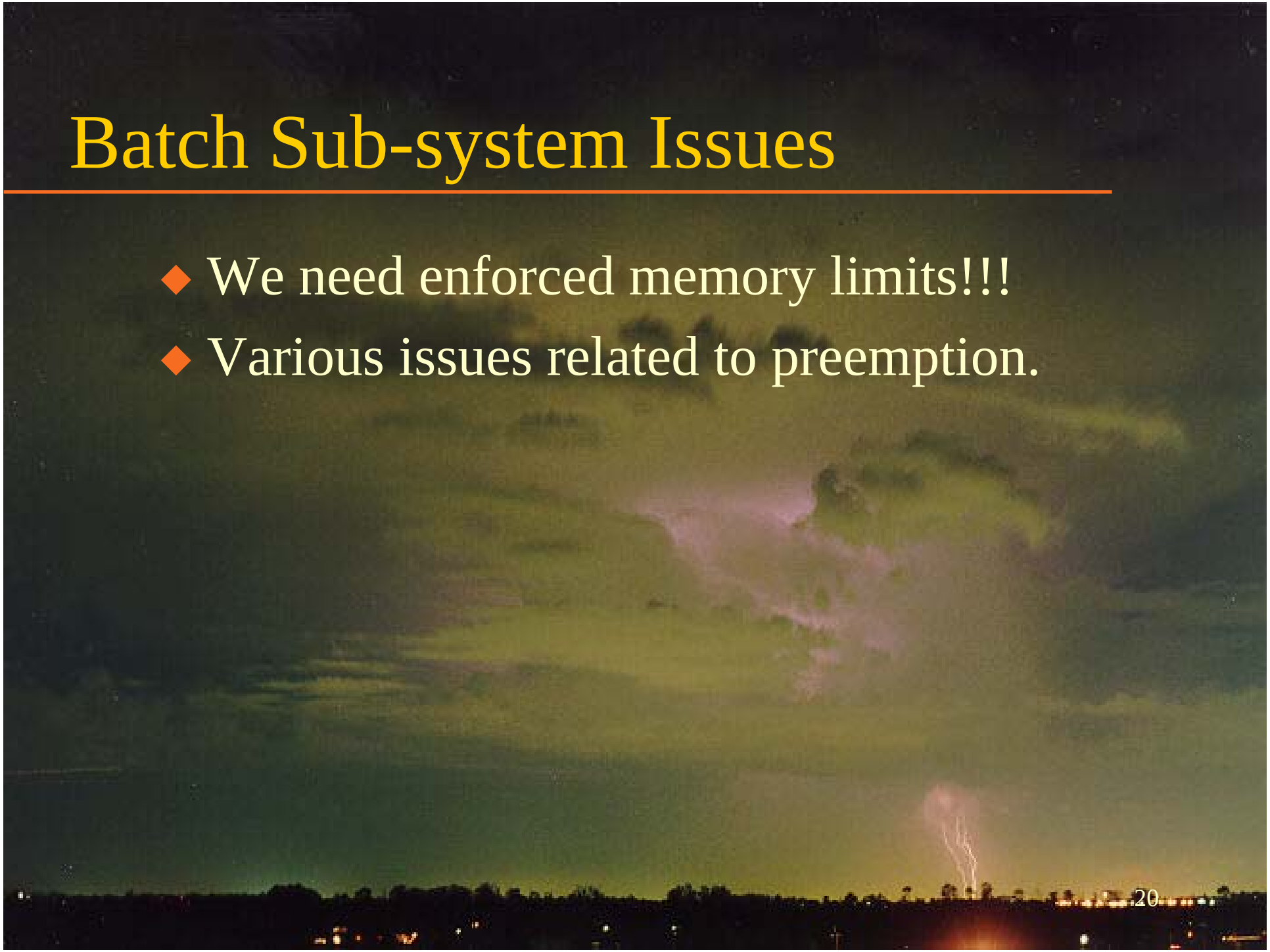


Legend
Production classe
Production daemon class
MSC classes
TEST class (azur-test)
MSC class (4 PE dedicated = 1 PE & < 15 min)

Color	Dedicated	#nodes	Name #
	VSD server	20	c1f0[12]p[1..7]m, c1f29p3m and c1f30p[123]m
	Interactive	2	c1f0[12]p8m
	< 8 CPU	3	c1f0[345]p1m
	Totalview License Server	1	c1f03p2m
	LL Negociator	1	c1f0[12]p8m
	LL Schedd	1	c1f0[12]p8m

Batch Sub-system Issues

- We need enforced memory limits!!!
- Various issues related to preemption.



System Usage

Monitoring

- Home grown tools
- Passive: log gathering and massaging
- Active: scripts pro-actively testing various items
- All syslogs redirected to the CWS
- How to cope with non-syslog logs? Still thinking…
- Heterogeneous solution.

Home grown stuff

- LL – SGE gateway
- Refurbished HPM toolkit
- Double buffered rcp.

On-going Support

- The systems must achieve a minimum of 99 % availability.
- Remedial maintenance 24/7: 30 minute response time between 8 A.M. and 5 P.M. on weekdays. One hour response outside above periods.
- Preventive maintenance or engineering changes: maximum of eight (8) hours a month, in blocks of time not exceeding two (2) or four (4) hours per day (subject to certains conditions) .
- Software support: Provision of emergency and non-emergency assistance.


SCIENTIFIC DIRECTIONS

Global System

Now:

- Uniform resolution of 100 km (400 X 200 X 28)
- 3D-Var at T108 on model levels, 6-hr cycle
- Use of raw radiances from AMSUA, AMSUB and GOES
- Use of MODIS satellite winds and profiler data.

- 2005:
  - Resolution to 35 km (800 X 600 X 58)
  - 4D-Var assimilation, 6-hr time window with 3 outer loops at full model resolution and inner loops at T108 (cpu equivalent of a 5-day forecast of full resolution model)
  - new datasets: QuikScat
- 2006+:
  - Additional datasets (AIRS, MSG, MTSAT, IASI, GIFTS, COSMIC)
  - Improved data assimilation

Regional System

Now:

- Variable resolution, uniform region at 15 km (575 X 641 X 58)
- 3D-Var assimilation on model levels at T108, 12-hr spin-up

2004:

- Resolution to 15 km in uniform region (576 X 641 X 58)
- Inclusion of AMSU-B and GOES data in assimilation cycle
- New datasets: profilers, MODIS winds

2005:

- Four model runs a day (instead of two)

2006+:

- LAM 4D-Var data assimilation
- Limited area model at 10 km resolution (800 X 800 X 60)
- Assimilation of Radar data

Ensemble Prediction System

Now:

- 16 members global system (300 X 150 X 28)
- Forecasts up to 10 days once a day at 00Z
- Optimal Interpolation assimilation system, 6-hr cycle,use of derived radiance data (Satems)

End 2004 (currently running in parallel):

- Ensemble Kalman Filter assimilation system, 6-hr cycle, use of raw radiances from AMSUA, AMSUB and GOES
- Forecasts extended to 15 days (instead of 10)

...Ensemble Prediction System

2005:

- Increased resolution to 100 km (400 X 200 X 58)
- Increased members to 32
- Additional datasets such as in global deterministic system
- Two forecast runs per day (12Z run added)

2007:

- Prototype regional ensemble
- 10 members LAM (500 X 500 X 58)
- No distinct data assimilation; initial and boundary conditions from global EPS

Mesoscale System

Now:

- Variable resolution, uniform region at 10 km (290 X 371 X 35)
  Two windows; no data assimilation

- Prototype Limited Area Model at 2.5 km (500 X 500 X 58)
  over one area

2006:

- Five Limited Area Model windows at 2.5 km (500 X 500 X 58)

2007+:

- 4D data assimilation

Coupled Models

Today

- In R&D: coupled atmosphere, ocean, ice, wave, hydrology, biology & chemistry
- In Production: storm surge, wave

2005

- Regional system coupled with ocean/ice model over Gulf of St Lawrence in operations.

Future

- Global coupled model for both prediction & data assimilation