Green computing

Computació sostenible i per a la sostenibilitat

Ivan Rodero Castro
Francesc Guim Bernat

PID_00191904

Universitat Oberta de Catalunya

Índex

Introducció..... 5

Objectius..... 6

1. Fonaments i conceptes bàsics..... 7
1.1. Mètriques 7
1.1.1. Mètriques per a equips individuals 8
1.1.2. Mètriques per a sistemes paral·lels 8
1.2. Caracterització del consum energètic 10
1.3. Llista Green500 12
1.4. Casos d'ús 13
1.4.1. BlueGene/Q 14
1.4.2. Projecte MontBlanc 15

2. Eficiència energètica..... 16
2.1. Fonts de potència en els circuits electrònics actuals 16
2.2. Mecanismes de gestió de l'energia 17
2.3. Gestió d'energia en el sistema operatiu 18
2.4. Gestió d'energia en centres de processament de dades 23
2.5. Gestió d'energia en computació d'altes prestacions 24
2.5.1. Anàlisi de la fase d'execució 24
2.5.2. Planificació i assignació de treballs 27

3. Computació sostenible i per a la sostenibilitat..... 29
3.1. Refrigeració natural 29
3.2. Cas d'ús: Parasol 30

4. Llista de lectures recomanades..... 32
4.1. Conceptes generals 32
4.2. Arquitectura i gestió de subsistemes 32
4.3. Entorns de computació d'altes prestacions 33
4.4. Centres de dades 34
4.5. Sostenibilitat 35

Bibliografia..... 37

Introducció

Aquest mòdul didàctic el plantegem com una breu introducció a la *green computing*, i fem un èmfasi especial en l'eficiència energètica per als sistemes d'altes prestacions, que servirà de referència, però que esperem que complementareu amb materials addicionals i lectures d'articles de recerca que hi estan relacionats.

Green computing és un terme molt ampli que cobreix des del desenvolupament d'arquitectures de baix consum fins a tècniques d'eficiència energètica com ara la utilització de fonts d'energia renovables.

Primer repassarem els fonaments bàsics d'aquest concepte, com ara les definicions i mètriques que hi estan associades. A continuació estudiarem diverses tècniques per a la millora de l'eficiència energètica en el context de diversos tipus de sistemes. També farem una breu introducció a alguns conceptes de sostenibilitat a partir de casos d'ús il·lustratius. Finalment, trobarem una llista de lectures recomanades dels temes més rellevants que tractem durant el mòdul.

En el primer mòdul didàctic de l'assignatura hem vist les principals motivacions per les quals l'eficiència energètica és clau en la computació d'altes prestacions. Cal recordar, però, que sense una millora d'ordres de magnitud en l'eficiència energètica no s'espera poder desenvolupar la pròxima generació de supercomputadors que han d'arribar a proporcionar rendiment d'exaflop en pocs anys, arran del creixent i desmesurat augment de la potència i les energies necessàries per a operar-los.

Objectius

Els materials didàctics d'aquest mòdul contenen les eines necessàries per a assolir els objectius següents:

1. Conèixer els fonaments i les mètriques utilitzades habitualment en el context de la *green computing* i l'eficiència energètica.

2. Conèixer les característiques i les tècniques principals de gestió de potència/energia elèctrica en general i per a sistemes d'altes prestacions.

3. Saber diferenciar entre *computació sostenible* i *computació per a la sostenibilitat* i conèixer-ne les característiques.

4. Conèixer les principals línies de recerca en l'àmbit de la *green computing* i l'eficiència energètica.

5. Adquirir una visió crítica d'articles de recerca relacionats amb la *green computing*.

1. Fonaments i conceptes bàsics

Green computing és un terme que normalment fa referència a la utilització eficient dels recursos informàtics, però és un concepte molt ampli. Tot i això, la motivació principal està relacionada amb la utilització de recursos informàtics per a minimitzar l'impacte ambiental, maximitzar la viabilitat econòmica i la sostenibilitat, i garantir les obligacions socials. La *green computing* també està molt relacionada amb altres moviments similars com ara la reducció de l'ús de materials ambientalment perillosos com els CFC, la promoció de l'ús de materials reciclables, la minimització de l'ús de components no biodegradables, i el foment de l'ús de recursos sostenibles.

Tot i que actualment hi ha molt d'interès en solucions que permetin utilitzar fons d'energia més verdes i sostenibles, nosaltres ens centrarem en l'eficiència energètica. Podem parlar d'eficiència energètica en diversos entorns, des de la producció (per exemple, amb tecnologia solar) i la distribució (en què es perd una gran quantitat d'energia) fins al consum (en què, a la pràctica, es desaprofita un gran percentatge de l'electricitat que es consumeix).

1.1. Mètriques

En aquest apartat veurem els fonaments més bàsics i mètriques relacionades amb la *green computing* i, concretament, amb l'eficiència energètica. Així doncs, a continuació veurem les definicions més bàsiques que farem servir en aquest mòdul i que cal tenir presents.

En el món de la física, l'**energia elèctrica** es pot definir com la forma d'energia que resulta de l'existència d'una diferència de potencial entre dos punts, que permet establir un corrent elèctric entre aquests punts quan són posats en contacte per mitjà d'un conductor elèctric per a obtenir treball. L'energia elèctrica es pot transformar en moltes altres formes d'energia, com ara l'energia tèrmica en el context dels circuits electrònics.

La **potència elèctrica** és la relació del pas de l'energia d'un flux per unitat de temps, és a dir, la quantitat d'energia alliberada o absorbida per un element en un temps determinat. Així doncs, l'energia és una funció d'integració de la potència al llarg del temps, i per això reduir la potència no ha de significar necessàriament una reducció d'energia.

Les unitats de mesura per a aquests dos conceptes en el sistema internacional són el joule (J) per a l'energia i el watt (W) per a la potència.

Podem definir molt genèricament l'**eficiència energètica** com l'obtenció d'un resultat que minimitza el consum d'energia o, complementàriament, totes les accions que tendeixen a reduir-ne el consum.

Les mètriques són essencials per a poder mesurar quantitativament, i per tant, per a avaluar l'eficiència del consum d'energia. Les mètriques formen la base per a la presa de decisions i, de fet, en els darrers anys s'han proposat diferents mètriques que actualment s'utilitzen.

A continuació en veurem de dos tipus diferents: per a equips individuals i per a sistemes paral·lels.

1.1.1. Mètriques per a equips individuals

La mètrica més bàsica d'eficiència energètica prové de la comunitat de disseny de circuits i és la fórmula ED^n. En aquesta fórmula, E és l'energia consumida durant l'execució d'una aplicació, D és el temps necessari per a completar-la i n és un paràmetre enter no negatiu que caracteritza l'equilibri entre E i D. Així doncs, aquesta mètrica combina energia i temps.

ED2P, que és el cas específic de ED^n quan $n = 2$, es la variant més utilitzada d'aquesta mètrica, especialment quan s'utilitzen tècniques de DVFS[1]. Amb aquesta mètrica es pot anul·lar la influència de l'escalat de la freqüència, ja que E és proporcional al quadrat de la freqüència, mentre que D^2 és proporcional a la inversa del quadrat de la freqüència. *ED2P* considera el rendiment i el consum d'energia, però no té en compte les necessitats dels diferents sistemes.

(1)*DVFS* és la sigla d'*escalat dinàmic de voltatge i freqüència* (*dynamic voltage and frequency scaling*).

Per a generalitzar aquesta mètrica, també es pot formular la mètrica *ED2P* amb pes (***weighted ED2P***), tal com es mostra a continuació:

$$Weighted \quad ED2P = E^{(1-\partial)} \times D^{2(1+\partial)}$$

En aquesta mètrica, $|\partial| \leq 1$ és un factor de pes determinat per les preferències de l'usuari. Aquesta mètrica intenta afavorir el rendiment quan $0 < \partial$, i en favor de l'energia quan $\partial < 0$. Si $\partial = 0$, el rendiment i l'energia són tractats de la mateixa manera, i té com a resultat la *ED2P* convencional.

1.1.2. Mètriques per a sistemes paral·lels

Quan el valor de n a ED^n és molt gran, es produeix un biaix a favor dels sistemes massivament paral·lels. De fet, la variant inversa de ED^n, és a dir, $1/ED^n$, representa $Rendiment^n/Potència$, o $Flops^n/W$, en què *Flops*, com ja sabem, fa re-

ferència al nombre d'operacions en coma flotant per segon. Quan un supercomputador té *s* processadors i cadascun d'aquests processadors proporciona *Fflops* amb *P* watts, la mètrica $Flops^n/W$ es pot reformular de la manera següent:

$$\frac{Flops^n}{W} = \frac{(s \cdot F)^n}{s \cdot P} = s^{n-1} \cdot \frac{F^n}{P}$$

En la llista Green500, que veurem més endavant, s'utilitza aquesta mètrica amb $n = 1$, ja que amb $n > 1$ el valor d'aquesta mètrica augmenta exponencialment amb el nombre de processadors *s*, i per tant, no seria fiable.

Vegeu també

Podeu veure la llista Green500 en el subapartat 1.3 d'aquest mòdul didàctic.

El ***TCO***[2] és el cost total de propietat d'un sistema informàtic i fa referència al cost total del sistema durant la seva vida, incloent-hi els costos d'adquisició, manteniment, energia consumida i eliminació. En els darrers anys, el cost de l'energia ha estat afectant el *TCO* gairebé al mateix nivell que la compra inicial. A més, hi ha costos de capital relacionats amb l'ocupació de l'espai que pesen considerablement en l'últim càlcul del *TCO*. L'eficiència energètica, la densitat, la refrigeració líquida, la fiabilitat, etc., són factors que contribueixen al baix cost de manteniment i, per tant, al *TCO*.

(2) *TCO* és la sigla de l'expressió anglesa *total cost of ownership*.

El ***PUE***[3] és sinònim d'eficàcia en l'ús de l'energia. Aquesta eficàcia es mesura a partir de la quantitat d'energia elèctrica que entra en un clúster o centre de dades (normalment s'utilitza més aviat en centres de dades) i que s'utilitza eficaçment per a fer càlculs, i per tant, és absorbida pels sistemes de tecnologia de la informació (*TI*[4]). El *PUE* es defineix de la manera següent:

(3) *PUE* és la sigla de l'expressió anglesa *power usage effectiveness*.

(4) *TI* és la sigla amb què es coneixen els sistemes de tecnologia de la informació.

$$PUE = \frac{\text{Potència total de la instal·lació}}{\text{Potència dels equips } TI}$$

Aquesta fórmula també es pot expressar de manera més detallada com es mostra a continuació:

$$PUE = \frac{\text{Refrigeració} + \text{Pèrdues de potència} + \text{Il·luminació} + TI}{TI}$$

El *PUE* teòric perfecte és igual a 1 i el *PUE* mitjà dels centres de dades és aproximadament de 2,13.

L' ***ERE***[5] fa referència a l'aprofitament de la calor generada pels sistemes informàtics. Aquesta tècnica s'anomena *energia tèrmica* i la reutilització implica que $PUE < 1$, que matemàticament no té sentit. L'*ERE* és una mètrica que pretén recollir aquesta idea d'energia total estalviada i es defineix de la manera següent:

(5) *ERE* és la sigla de l'expressió anglesa *energy reuse effectiveness*.

$$ERE = \frac{\text{Energia total} - \text{Energia reutilitzada}}{\text{Energia consumida per } TI}$$

El ***CUE***[6] representa l'eficàcia en l'ús de carboni. Mesura el total d'emissions de CO_2 causades pel centre de dades dividit per l'energia de la càrrega del sistema, que és l'energia consumida pels servidors. La fórmula es pot expressar de la manera següent:

(6)*CUE* és la sigla de l'expressió anglesa *carbon usage effectiveness*.

$$CUE = \frac{CO_2 \text{ emès } (KgCO_2eq)}{\text{Unitat d'energia (Kwh)}} \times \frac{\text{Energia total del sistema}}{\text{Energia consumida per } TI}$$

És a dir:

$$CUE = CEF \times PUE$$

En aquesta darrera expressió, *CEF* és el factor d'emissió de carboni (kgCO2eq/kWh) del sistema, d'acord amb les dades publicades pel govern de la regió d'operació. El *CEF* depèn de la barreja de producció d'energia que en última instància alimenta el centre de dades. Aquest indicador pot ser molt baix, com, per exemple, en el cas d'un centre de dades que funciona a partir de l'electricitat generada per una central hidroelèctrica. En realitat, el *CEF* canvia de país a país i s'actualitza anualment. Als EUA, canvia fins i tot d'estat a estat i la mitjana dels EUA és 0,59 kgCO2eq/kWh.

A banda d'aquestes mètriques, n'hi ha dues més que s'acostumen a utilitzar per a mesurar l'eficiència energètica en relació amb la productivitat. La primera és la productivitat de la tecnologia de la informació per watt (*IT-PEW*[7]). La segona és l'índex d'eficiència energètica i productivitat (*DC-EEP*[8]). Aquestes mètriques es defineixen amb les equacions següents:

(7)*IT-PEW* correspon a l'expressió anglesa *IT productivity per embedded watt*.

(8)*DC-EEP* correspon a l'expressió anglesa *data center energy efficiency and productivity*.

$$IT - PEW = \text{Productivitat} / Embedded \quad watt$$

$$DC\text{-}EEP \quad Index \; = \; PEW/PUE \; = \; \text{Productivitat / Potència total cluster}$$

Aquí, la *Productivitat* és la producció de servei del sistema i el terme *Embedded watt* és la potència dels sistemes de tecnologia d'informació. Així doncs, mentre que *IT-PEW* indica l'eficiència energètica dels equips informàtics, *DC-EEP index* indica la corresponent de tot el clúster o centre de dades.

També hi ha altres mètriques més relacionades amb els centres de dades a Internet com ara rendiment per watt, rendiment per *TCO* i rendiment per cost de l'energia i refrigeració.

1.2. Caracterització del consum energètic

Per a fer una gestió eficient de l'energia, cal conèixer i poder caracteritzar diferents patrons d'ús de l'energia del sistema. En altres paraules, necessitem saber on i quan s'ha consumit l'energia i qui és responsable d'usar-la. La construcció d'un perfil d'aquestes característiques s'anomena *perfil de consum d'energia*.

A partir d'aquests **perfils de consum d'energia** s'han pogut desenvolupar diferents tècniques d'anàlisi com ara les basades en simulacions, la modelització analítica, la mesura directa del consum d'energia, l'anàlisi de consum d'energia basat en el monitoratge, o tècniques de mostratge de la potència i programari d'instrumentació, que passem a detallar a continuació:

1) En les **tècniques basades en simulacions**, les característiques de consum d'energia estan integrades en un simulador, el qual està basat en les característiques de potència derivades de mesures d'una mostra. Aquests simuladors estimen el rendiment i el consum d'energia mitjançant el seguiment de l'execució de les aplicacions. Les tècniques de simulació s'han utilitzat per a diversos tipus d'entorns i de components com ara el processador, la memòria, el disc dur o fins i tot màquines senceres. El mecanisme de simulació és útil per a caracteritzar l'activitat i la potència, però hi pot haver un desajust entre la simulació i els sistemes reals a causa de la inexactitud dels models de simulació. A més, les simulacions poden arribar a tardar força temps a executar-se.

2) Les **tècniques analítiques de modelització** abstreuen el consum d'energia mitjançant funcions analítiques sobre un conjunt de paràmetres. El consum d'energia s'estima a partir de correlacions entre la potència elèctrica i les variables proporcionades. Aquest tipus pot servir per a modelitzar, per exemple, l'emmagatzematge compost per un conjunt de discos redundants. En aquest cas, els paràmetres relacionats amb l'energia són (a) la potència necessària per a cada disc en mode actiu o llest, i (b) l'energia i el temps necessaris perquè un disc es pugui engegar i apagar. Tenint en compte els paràmetres dels discos i les seves altres característiques, el model prediu el comportament de consum elèctric dels sistemes d'emmagatzematge, la política de gestió de disc i la càrrega de treball que hi està associada. Es poden utilitzar equacions lineals i també es poden utilitzar qüestions de gestió tèrmica dels sistemes. Un exemple és el desenvolupament de models que emulen la temperatura a partir de les característiques dels components electrònics. La modelització analítica té l'avantatge de ser una solució basada en programari; tot i això, l'exactitud de la solució depèn de la granularitat del model.

3) La **mesura en temps real del consum d'energia** és una solució directa al problema de la caracterització del consum elèctric. Per a dur-la a terme, cal emprar eines de mesura addicionals, com ara multímetres, que mesurin i enregistrin el consum d'energia en temps d'execució. Un tema clau és la selecció de l'aplicació apropiada que serveixi com a referència[9] i que sigui representativa. No obstant això, aquest enfocament no és pràctic per a la majoria dels sistemes, sobretot per a sistemes d'una certa escala en què caldria milers de multímetres, la qual cosa no sempre és viable des del punt de vista econòmic.

(9)En anglès, *benchmark*.

4) El **monitoratge basat en l'anàlisi de consum de potència** es fa a partir dels comptadors de rendiment (PMC) que hi ha integrats en els processadors actuals. A partir d'aquests comptadors es poden extreure perfils de consum d'energia prou acurats, els quals s'han utilitzat per a la gestió d'energia tèrmi-

ca. No obstant això, l'exactitud dels comptadors, i per tant la caracterització del tipus de perfil, està limitat pels tipus d'esdeveniments que poden ser monitorats. Per exemple, les mesures relacionades amb operacions que es fan fora del processador són habitualment molt menys precises.

5) Les **tècniques basades en el mostratge** intenten buscar correlacions entre el comportament del programari/sistema i el consum d'energia del sistema. Un exemple d'aquesta tècnica consisteix a mesurar periòdicament el consum d'energia juntament amb el comptador de programa (PC) i el procés ID (PID), i tornar a vincular el consum d'energia de la mostra als PID (és a dir, els processos) i els PC (és a dir, les fases d'execució). Les tècniques de mostratge no cal que depenguin del maquinari, però en aquest cas és més difícil de determinar quins components utilitzen la majoria de l'energia, i pot ser que el mostratge no sigui prou efectiu.

6) Les **tècniques basades en instrumentació** afegeixen fragments addicionals en el codi de les aplicacions amb l'objectiu de recollir informació sobre l'alimentació i el context de l'execució del programa. Alguns exemples d'aquesta tècnica són: inserir codi de perfils en el codi assemblador que mesura el consum d'energia o utilitzar eines d'instrumentació per a inserir dinàmicament rutines de generació de perfils en el codi binari. La instrumentació es pot utilitzar per a construir un model d'energia de gra fi, però requereix un compilador de propòsit especial o suport d'eines addicionals per a modificar el codi binari o font.

1.3. Llista Green500

La llista dels Green500 agrupa els supercomputadors ordenats per eficiència energètica. L'objectiu d'aquesta llista és deixar patents els esforços per millorar la capacitat de computació sense penalitzar el consum energètic.

A diferència dels Top500, en què la mètrica principal és el rendiment (mesurat en megaflops), la llista dels Green500 se centra en l'eficiència energètica (mesurada en megaflops per watt).

Vegeu també

Podeu veure la llista dels Top500 en el subapartat 4.5.2 del mòdul "Introducció a la computació d'altes prestacions" d'aquesta assignatura.

La taula 1 mostra la llista dels deu supercomputadors energèticament més eficients. En general, veiem que no coincideix gens amb la llista dels Top500, excepte en el cas de Titan, que és el tercer en la llista dels Green500, i és el computador més potent en la llista dels Top500 de novembre del 2012.

De la mateixa manera que en els Top500, la llista s'elabora a partir de l'execució del *bechmark* HPL[10]. La diferència resideix en el fet que hi ha unes normes per al monitoratge de la potència del sistema que inclouen utilitzar multímetres quan el sistema executa HPL en tot el sistema i aquest ha arribat a la temperatura operativa (almenys després de 15 minuts).

(10)*HPL* és la sigla de l'expressió anglesa *high performance linpack*.

Taula 1. Llista dels deu supercomputadors més eficients energèticament de la llista dels Green500 de novembre del 2012

#	MFlops/W	Institució	Computador	Potència (kW)
1	2.499,44	National Institute for Computational Sciences / University of Tennessee	Beacon - Appro GreenBlade GB824M, Xeon E5-2670 8C 2.600 GHz, Infiniband FDR, Intel Xeon Phi 5110P	44,89
2	2.351,10	King Abdulaziz City for Science and Technology	SANAM - Adtech ESC4000/FDR G2, Xeon E5-2650 8C 2.000GHz, Infiniband FDR, AMD FirePro S10000	179,15
3	2.142,77	DOE/SC/Oak Ridge National Laboratory	Titan - Cray XK7, Opteron 6274 16C 2.200 GHz, Cray Gemini interconnect, NVIDIA K20x	8.209,00
4	2.121,71	Swiss Scientific Computing Center (CSCS)	Todi - Cray XK7, Opteron 6272 16C 2.100 GHz, Cray Gemini interconnect, NVIDIA Tesla K20 Kepler	129,00
5	2.102,12	Forschungszentrum Juelich (FZJ)	JUQUEEN - BlueGene/Q, Power BQC 16C 1.600 GHz, Custom Interconnect	1.970,00
6	2.101,39	Southern Ontario Smart Computing Innovation Consortium / University of Toronto	BGQdev - BlueGene/Q, Power BQC 16C 1.600 GHz, Custom Interconnect	41,09
7	2.101,39	DOE/NNSA/LLNL	rzuseq - BlueGene/Q, Power BQC 16C 1,60 GHz, Custom	41,09
8	2.101,39	IBM Thomas J. Watson Research Center	BlueGene/Q, Power BQC 16C 1,60 GHz, Custom	41,09
9	2.101,12	IBM Thomas J. Watson Research Center	BlueGene/Q, Power BQC 16C 1,60 GHz, Custom	82,19
10	2.101,12	École Polytechnique Fédérale de Lausanne	CADMOS BG/Q - BlueGene/Q, Power BQC 16C 1.600 GHz, Custom Interconnect	82,19

Font: The Green 500

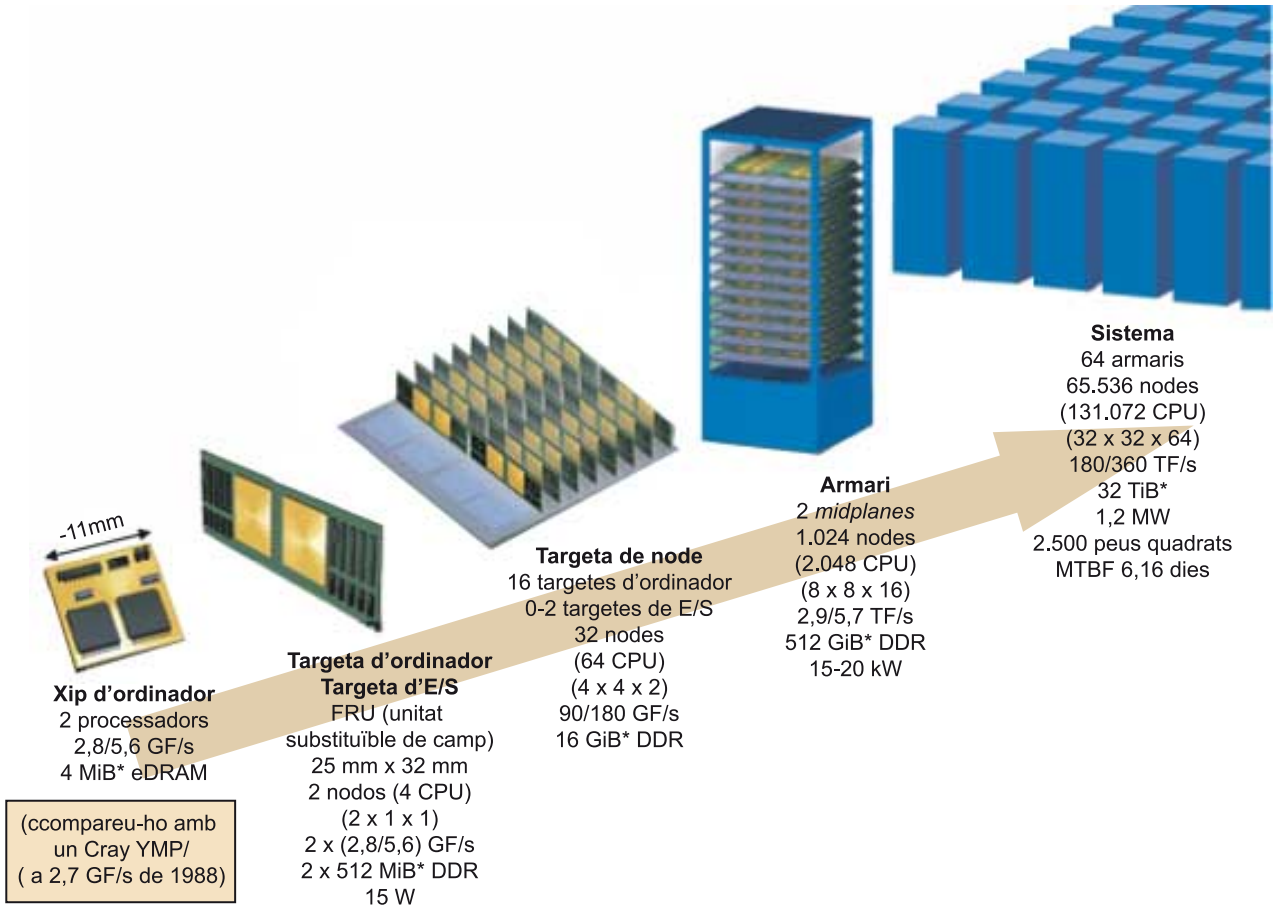
1.4. Casos d'ús

En aquest subapartat utilitzarem dos casos d'ús específics de sistemes que es van dissenyar amb l'objectiu de ser molt eficients energèticament tot i que tenen característiques molt diferents però son casos prou representatius.

1.4.1. BlueGene/Q

Tal com es pot observar en la taula 1, la meitat dels deu supercomputadors més eficients energèticament són BlueGene/Q, que va dissenyar IBM i precedeix el BlueGene/L i el BlueGene/P. La idea fonamental d'aquest disseny és obtenir un disseny molt compacte, tal com es pot observar en la figura 1, que permeti obtenir un alt rendiment (petaflops) amb un consum reduït.

Figura 1. Diagrama de la jerarquia de l'arquitectura BlueGene (en concret, es mostra la de BlueGene/P)



Font: IBM

A més, el BlueGene/Q disposa d'un programari per a gestionar l'energia – l'IBM Systems Director Active Energy Director, que és l'eix central de gestió de l'energia. Aquest programari mesura, gestiona i controla la potència i l'ús d'energia tèrmica i també s'integra en la infraestructura i els paquets de gestió empresarial. Per a aconseguir escalabilitat fins a milions de nuclis per a assolir rendiment d'exaflop, la gestió de recursos subjacent a l'arquitectura de programari ha de proporcionar un mecanisme flexible per a donar suport a la gran quantitat d'aplicacions de diferents característiques i tipus de càrrega de treball que s'executen en aquesta plataforma. LoadLeveler, el planificador d'IBM, per a fer la seva tasca, conscient de l'energia, estableix la freqüència del processador de manera òptima en el conjunt de nodes en què s'executa un treball i també la freqüència dels nodes perquè consumeixin el mínim possible d'energia quan aquests nodes no tenen tasques.

1.4.2. Projecte MontBlanc

El projecte MontBlanc és un projecte europeu, inicialment de tres anys de durada, que té per objectiu dissenyar un supercomputador basat en la tecnologia de baix consum que s'utilitza actualment en tauletes i en telefonia mòbil. Amb aquesta tecnologia, els investigadors pretenen dissenyar un ordinador de prestacions idèntiques, però amb un consum energètic entre quatre i deu vegades més baix.

MontBlanc integra tecnologia de baix consum d'ARM i acceleradors dissenyats per a dispositius mòbils amb l'objectiu de desenvolupar una nova classe de supercomputador igual de potent però molt més eficient en l'ús d'energia.

Si considerem que la potència elèctrica disponible marca el límit del rendiment d'un supercomputador, aquesta tecnologia ens permetria disposar de sistemes entre quatre i deu vegades més potents. Tot i això, hi ha reptes molt importants que cal solucionar, com ara la programabilitat, la jerarquia de memòria i la xarxa d'interconnexió d'aquest tipus d'arquitectura, ja que haurà de donar suport a un nivell de paral·lelisme molt superior al d'altres tipus d'arquitectures basades en processadors més potents i amb més requeriments de potència.

En aquest projecte, finançat per la UE, hi ha empreses europees líders en el sector tecnològic, com Bull i ARM, i els centres de supercomputació de més pes a Europa.

2. Eficiència energètica

En aquest apartat ens centrarem en l'eficiència energètica i les tècniques de gestió de potència i energia elèctrica tant en sistemes d'altes prestacions com en sistemes computacionals en general, com són els centres de dades a Internet.

Tot i que veurem tècniques d'eficiència energètica per a diferents tipus de dispositius i en diferents àmbits, fins i tot en l'àmbit de programari, primer estudiarem la font principal de potència dels circuits electrònics i després ens centrarem en les tècniques que ens permeten gestionar-ne l'eficiència.

2.1. Fonts de potència en els circuits electrònics actuals

El consum d'energia elèctrica en un circuit electrònic prové de dues fonts: el consum estàtic i el consum dinàmic.

El **consum estàtic** és el que es produeix a causa dels corrents de fuga[11] que hi ha en els transistors.

[11] En anglès, *leakage*.

El consum estàtic és inherent al circuit, fins i tot quan el circuit està inactiu. Amb l'avenç de la tecnologia, aquest component de la potència és cada vegada més important. El valor d'aquest consum depèn de les característiques de la tecnologia que s'empra, el nombre de transistors i la temperatura de funcionament del circuit.

La potència estàtica $P_{Estàtica}$ es defineix com el producte del voltatge de la font d'alimentació V_s pel corrent estàtic del circuit i_0, tal com indica l'equació següent:

$$P_{Estàtica} = \sum_{1}^{n} i_0 \cdot V_s; \qquad i_0 = i_s\left(e^{\frac{qV_{Díode}}{KT}} - 1\right)$$

Aquí, i_s és el corrent invers de saturació, o corrent de fuga dels díodes, $V_{Díode}$ és el voltatge del díode, q és la unitat de càrrega, K és la constant de Boltzmann i T és la temperatura.

El **consum dinàmic** es produeix per la càrrega i descàrrega de la capacitat dels transistors i les connexions, i depèn de l'activitat del circuit.

És a dir, passa únicament durant les transicions, quan les portes commuten. Per tant, és proporcional a la freqüència de commutació, i com més gran sigui el nombre de commutacions, més gran serà la potència dinàmica. L'equació següent mostra la potència dinàmica:

$$P_{Dinàmica} = a \cdot C \cdot f \cdot V_s^2$$

En aquesta expressió, *a* és l'activitat de commutació, *C* és la capacitat en cada node que commuta, *f* és la freqüència de rellotge i V_s és el valor del potencial d'alimentació.

La potència dinàmica té dos components, com mostra l'equació següent: la potència de commutació[12] i la potència de càrrega[13]. La de commutació és deguda als corrents que van de la font d'alimentació a terra quan el transistor canvia d'estat, mentre que la de càrrega es deu al corrent necessari per a carregar les capacitats dels elements connectats a la sortida.

[12] En anglès, *crowbar*.

[13] En anglès, *load*.

$$P_{Dinàmica} = P_{Commutació} + P_{Càrrega}$$

2.2. Mecanismes de gestió de l'energia

Hi ha dos tipus bàsics de mecanismes de gestió de l'energia, l'escalat dinàmic de velocitat i l'adormiment dinàmic de recursos:

1) L'escalat dinàmic de velocitat[14] és un mecanisme que canvia dinàmicament l'estat de funcionament del component en qüestió per tal de reduir la potència, és a dir, s'alenteix per reduir el consum d'energia i s'accelera quan és necessari, però a costa de més consum d'energia.

[14] En anglès, *dynamic speed scaling*.

Alguns exemples d'escalat dinàmic de velocitat

Un exemple típic és l'escalat dinàmic de voltatge i freqüència (DVFS, de l'anglès *dynamic voltage and frequency scaling*). En aquest cas, la reducció en el consum d'energia es fa mitjançant la reducció del voltatge d'alimentació o la freqüència de rellotge. La majoria dels processadors actuals són compatibles amb aquest mecanisme. Alguns exemples d'aquest escalat són els processadors Intel Xeon i els processadors o coprocessadors AMD.

La regulació tèrmica n'és un altre exemple. En aquest cas el que es controla és la temperatura del processador. De la mateixa manera que abans, això es fa mitjançant la modulació del cicle de treball del rellotge del processador o la reducció de la freqüència de funcionament i el voltatge del processador.

Uns altres exemples de DSS inclouen memòries multifreqüència, en què es pot escalar la freqüència de treball dinàmicament i, per tant, la velocitat d'accés a les dades, i discos amb múltiples velocitats.

No obstant això, en tots aquests mètodes, la transició entre diferents estats de funcionament consumeix energia addicional i causa sobrecàrrega en la latència.

2) L' **adormiment dinàmic de recursos**[15] és un mecanisme que adorm (o hiberna) components dinàmicament per tal d'estalviar energia i els desperta quan són necessaris. Cada component pot estar en un estat actiu, en un dels estats de son, o bé en estat d'apagada. Tal com veurem més endavant, en

[15] En anglès, *dynamic resource sleeping*.

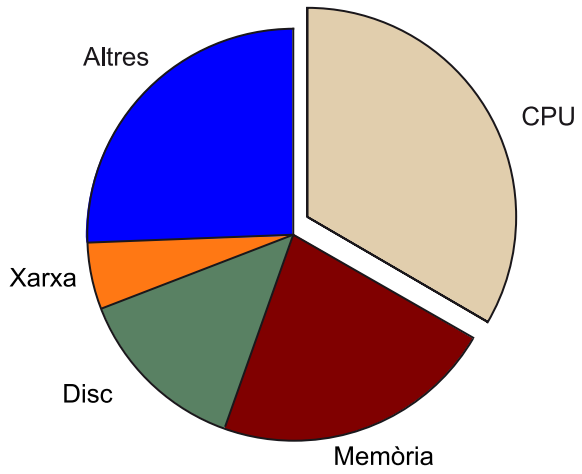
l'estàndard de la indústria ACPI, l'estat actiu es denota C0 i els estats de son, C1, C2 ... Cn. Cadascun dels estats de son consumeixen menys energia que l'estat C0 sense activitat. Com més profund és l'estat de son del processador, menys potència consumeix, però més energia es necessita per a despertar-lo. Els controladors de memòria també poden canviar la gestió dinàmica de la potència i els discos també poden funcionar amb els estats actiu, llest i en espera. De fet, tot l'equip es pot administrar també com si fos un component que pot estar en els estats actiu, suspès, hibernació o apagat, però les transicions entre estats consumeixen energia i tarden un cert temps.

Tot i que el processador és el principal subsistema que dissipa potència elèctrica en els computadors actuals, cada cop més els altres subsistemes estan augmentant la demanda de potència. La figura 2 mostra una possible distribució de la potència dissipada pels diferents subsistemes d'un computador estàndard.

Hi ha força tècniques que intenten fer un ús eficient d'aquests subsistemes, tal com veurem en els subapartats següents, entre les quals destaquen les arquitectures de baixa potència, DVFS[16], apagament de subsistemes (fins i tot de tot el computador), utilització d'estats de baix consum o planificació eficient de la càrrega.

(16) *DVFS* és la sigla de *dynamic voltage and frequency scaling.*

Figura 2. Distribució de potència elèctrica d'un computador per subsistema



2.3. Gestió d'energia en el sistema operatiu

En aquest subapartat tractarem d'una de les tècniques de gestió de l'energia més estesa i que des de fa temps funciona amb els sistemes operatius de manera transparent per a l'usuari. Aquesta tècnica és complementària d'altres tècniques que exposem en aquest mòdul didàctic.

L'especificació de les **interfícies avançades de configuració i energia** (ACPI[17]) va ser desenvolupada per a establir interfícies comunes a la indústria que permetessin gestionar energia i configurar-ne en dispositius que deleguen aquesta responsabilitat al sistema operatiu. En la indústria, això últim és conegut amb la sigla OSPM[18].

(17)*ACPI* és la sigla de l'expressió anglesa *advanced configuration and power interface.*

(18)OSPM és la sigla d'*operating system-directed configuration and power management*.

L'ACPI és la consolidació de diversos intents i aproximacions anteriors que tracta de respondre a les seves mateixes necessitats des d'un punt de vista menys heterogeni i més flexible. L'ACPI agrupa i substitueix rutines de codi localitzades a la BIOS en una especificació d'interfície ben definida, encara que extensa i molt complexa. L'ACPI també proporciona els mecanismes necessaris per a fer una transició ordenada entre el maquinari més antic i el més recent.

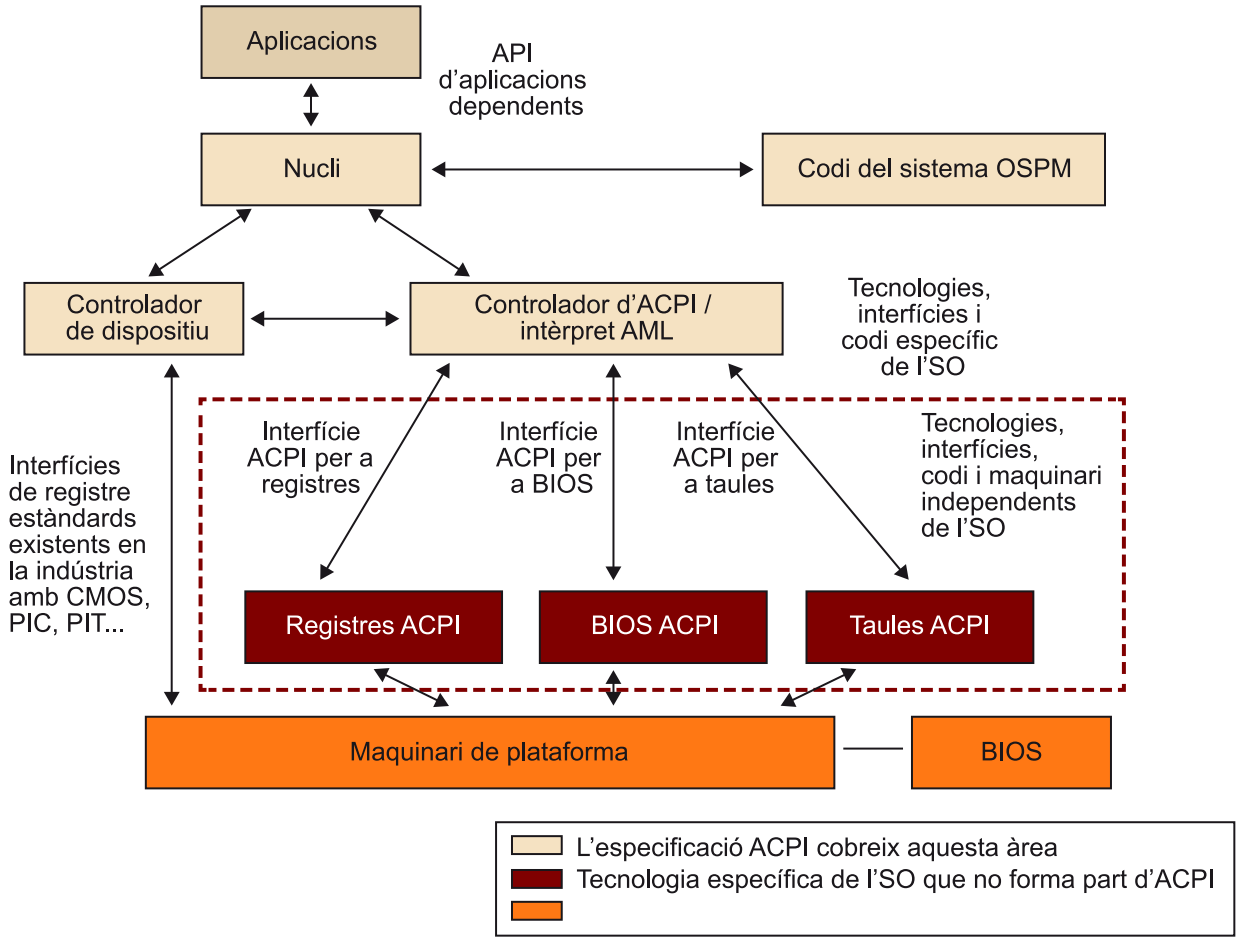
Les interfícies i el concepte *OSPM* que es recullen en l'especificació són transversals a les diferents implementacions d'ordinadors genèriques que hi ha a la indústria. Per exemple, ordinadors personals, estacions de treball, telèfons mòbils o servidors són implementacions que es poden beneficiar d'aquesta especificació OSPM/ACPI. Cal dir que els sistemes operatius més difosos i emprats en el mercat d'ordinadors (per exemple, Microsoft Windows, GNU/Linux i BSD) disposen per defecte de suport per a ACPI habilitat en major o menor grau.

L'especificació fa servir el concepte de *conservació de l'energia* per mitjà de la transició de dispositius en estats de baix consum, o fins i tot de no consum quan no fan treball útil.

L'ACPI descriu les interfícies de maquinari, programari i les estructures de dades que, un cop implementades, activen el suport per a dur a terme l'OSPM i permeten fer la gestió d'energia des del mateix sistema operatiu (SO) emprant una interfície abstracta entre el sistema operatiu i el maquinari. L'ACPI també inclou la semàntica d'aquestes interfícies. La figura 3 mostra els components de programari i maquinari més rellevants en OSPM/ACPI.

Com s'observa en la figura 3, l'especificació descriu les interfícies entre components, els continguts de les taules de descripció de sistema ACPI i la semàntica relacionada de la resta de components.

Figura 3. Esquema global del sistema OSPM/ACPI



Font: adaptat de l'especificació ACPI

Les taules de descripció de sistema ACPI, que descriuen un maquinari de plataforma concret, són el cor de la implementació ACPI al mateix temps que el microprogramari de sistema ACPI porta a terme, entre les responsabilitats que té, el subministrament de taules ACPI (que són independents de la tecnologia) i no així el d'una interfície nativa, que seria menys flexible.

Les **responsabilitats funcionals de l'OSPM** són, per tant, disposar del control d'accés directe i exclusiu sobre les funcions de configuració i gestió d'energia del maquinari.

D'aquesta manera, durant la inicialització, l'OSPM és responsable de gestionar els esdeveniments de configuració generats pel maquinari i de controlar el consum d'energia, el rendiment i l'estat tèrmic del sistema tenint en compte sempre les preferències d'usuari, les peticions pel que fa a l'aplicació, i els objectius d'usabilitat i qualitat de servei.

Les àrees funcionals que permeten a l'OSPM fer aquestes funcions les descrivim a continuació:

- *System power management*. L'ACPI defineix mecanismes per permetre que l'ordinador passi a estats de baix consum. Això es pot dur a terme a escala de sistema o de dispositiu.

- *Device power management*. Les taules ACPI descriuen el maquinari i els seus estats d'energia i permeten posar un dispositiu en diferents estats de baix consum.

- *Processor power management*. Mentre el sistema operatiu està en repòs, l'ACPI permet posar el processador en estats de baix consum.

- *Device and processor performance management*. Mentre el sistema està actiu, l'OSPM permet fer transicions entre estats per a dispositius i processadors amb la intenció d'obtenir el compromís desitjat entre rendiment i conservació de l'energia.

- *Configuration / Plug and play*. L'ACPI especifica informació que s'utilitza per a enumerar i configurar el maquinari.

- *System events*. L'ACPI defineix mecanismes molt flexibles per encaminar esdeveniments a la lògica de cada dispositiu en maquinari.

- *Battery management*. La política de gestió de bateria mou des de la BIOS cap a l'ACPI. Les bateries compatibles amb l'ACPI disposen d'una petita interfície definida per a mètodes de control.

- *Thermal management*. L'ACPI funciona amb gestió tèrmica de manera que proporciona un model escalable a fabricants que els permeti definir zones, indicadors i mètodes de control per a una gestió tèrmica correcta.

- *Embedded controller / SMBus controller*. L'ACPI defineix un maquinari estàndard i un programari de comunicacions que fa d'interfície entre controladors del sistema operatiu i un controlador SMBus que permet als fabricants proporcionar funcionalitats perquè les puguin utilitzar el sistema operatiu i les aplicacions.

L'ACPI reconeix cinc estats globals (*global states*), i un nombre d'estats que succeeixen en algun dels cinc estats globals esmentats, tal com es mostra en la figura 4. Aquests estats es poden classificar en les categories següents:

1) *Global states* (G0-G3). Són estats globals amb la definició següent:

- G0: actiu (*working state*)
- G1: en espera (*sleeping state*)
- G2: mode inactiu (*soft-off state*)
- G3: apagament mecànic (*mechanical-off state*)

Hi ha també un cinquè estat global, no considerat dins d'aquest grup, que és anomenat *legacy*. Aquest darrer estat representa l'estat del sistema quan no funciona amb el model ACPI.

2) *C-states*. Tenen lloc en el context de G0 (*global working state*). C0, un dels *C-states* (*CPU-state*), fa referència a l'estat d'execució. En canvi, si volem estalviar energia quan el processador està desocupat, és preferible estats amb un valor més alt. De fet, cap instrucció no s'executa en C1, C2 i C3. Cal dir que l'ACPI substitueix el típic bucle d'*idle* per defecte per poder entrar en C1, C2 i C3.

3) *P-states*. En el context de G0 (*global working state*) i C0 (*CPU executing state*), tenen lloc els *P-states* (*performance states*). Aquests estats serveixen per a modular la freqüència i el voltatge del processador quan està executant instruccions. Són estats molt efectius i tenen un subsistema propi en el domini de la gestió d'energia del nucli Linux (*cpufreq*). Tenen un paper notable en la tècnica de *throttling*, en què el nucli tracta d'obtenir un compromís entre la freqüència, la potència i el rendiment desitjat.

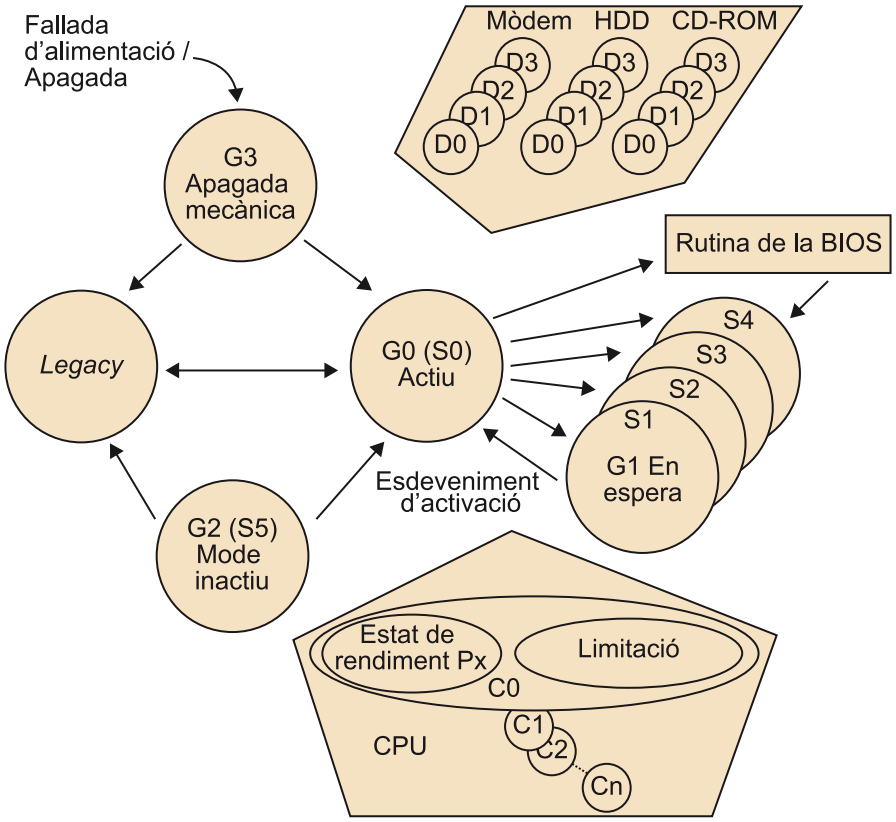
4) *Sleep states*. L'ACPI disposa dels estats S0-S5 amb el significat següent:

- S0 és un estat en què no està en espera (*non-sleep state*).
- S1 és l'estat d'espera (*stand-by*), en el qual el processador està aturat i la pantalla apagada.
- S2 no s'utilitza.
- S3 és suspendre el sistema operatiu i mantenir-lo a la memòria RAM.
- S4 és per a hibernar la imatge del sistema operatiu a disc.
- S5 és per a mode inactiu (*soft-power off state*).

4) *Device states* (D0-D3). L'ACPI defineix estats de baix consum per als dispositius de la manera següent:

- D0 està encès.
- D3 està apagat.
- D1 i D2 són estats intermedis.

Figura 4. Estats d'energia Global System i transicions en ACPI global del sistema OSPM/ACPI



Font: adaptat de l'especificació ACPI

2.4. Gestió d'energia en centres de processament de dades

En centres de dades comercials i a Internet, hi ha diverses tècniques de gestió d'energia. Algunes d'aquestes tècniques es basen en l'engegament i apagament de servidors de manera que es conserva energia alhora que es respecta la qualitat de servei, la qual es defineix a partir dels SLA[19]. Hi ha altres tècniques que es basen en l'ús de DVFS de manera que es pot millorar l'eficiència energètica sense penalitzar el rendiment, com, per exemple, en la manera de limitar la potència del sistema. En el darrer cas, normalment es fa amb un controlador. Fins i tot s'utilitzen sensors de temperatura, flux d'aire, etc., per a fer una gestió tèrmica eficient, i per tant, energètica.

(19) *SLA* és la sigla de l'expressió anglesa *service level agreements*.

També en centres de dades virtualitzats hi podem trobar tècniques específiques com ara provisió, planificació i consolidació de màquines virtuals tenint en compte l'eficiència energètica (per exemple, a partir dels perfils de les aplicacions), i també aspectes tèrmics com, per exemple, el control dels sistemes de refrigeració o emplaçament de màquines virtuals intel·ligents.

Atès que aquest tipus de tècniques es cobreixen en les lectures recomanades en l'últim apartat d'aquest mòdul didàctic, no les desenvoluparem aquí, ja que el focus d'aquest mòdul didàctic és la computació d'altes prestacions.

2.5. Gestió d'energia en computació d'altes prestacions

Un sistema d'altes prestacions típic es pot dividir en tres subsistemes: els nodes frontals (*front-end*), els nodes computacionals i els nodes d'emmagatzematge:

a) Els **nodes frontals** ofereixen una interfície a l'usuari perquè pugui accedir al sistema i així enviar treballs al sistema i monitorar-ne l'estat.

b) Els **nodes de computació** proporcionen recursos de computació d'alt rendiment per processar els treballs.

c) Els **nodes d'emmagatzematge** proporcionen una capacitat d'emmagatzematge massiva. Aquests tres subsistemes estan integrats en un sistema fortament acoblat per una xarxa d'interconnexió d'alta velocitat.

A diferència dels clústers comercials, les aplicacions científiques normalment són altament acoblades[20], que generalment impliquen molts nodes, fins i tot un nombre massiu, que treballen de manera coordinada. La comunicació i la sincronització s'entrellacen amb el càlcul, i el temps d'execució és relativament llarg. En aquest context, la CPU domina el consum d'energia en el sistema d'altes prestacions (vegeu la figura 2). Aquestes característiques dels sistemes d'altes prestacions imposen certs reptes i oportunitats per a la gestió de la seva energia, que es poden classificar en dues categories bàsiques:

[20] En anglès, *tight-couple*.

- Anàlisi de la fase d'execució.
- Planificació i assignació de treballs.

2.5.1. Anàlisi de la fase d'execució

En termes generals, les aplicacions paral·leles consisteixen en computació amb la CPU, accés a memòria, entrada/sortida i comunicació/sincronització. En un supercomputador, normalment l'entrada/sortida habitual s'implementa per la interfície de comunicació per a accedir a un subsistema d'emmagatzematge (sistema de fitxers distribuïts). Per tant, aquest cas serà un cas especial de comunicació.

Així doncs, podem dir que hi ha tres tipus de fases durant l'execució de l'aplicació: basades en CPU, en memòria i en comunicacions. Quan una fase està basada en un component els altres normalment es poden posar en estats de baix consum sense degradar significativament el rendiment de l'execució de l'aplicació. S'han desenvolupat una sèrie de tècniques per a poder aprofitar els recursos a la vegada que es fa una gestió eficient de l'energia. A continuació estudiarem algunes de les tècniques més típiques que se centren en cadascun dels subsistemes que acabem de comentar.

Anàlisi de l'activitat de la CPU

Una de les tècniques més típiques per a la gestió de potència és utilitzar DVFS basant-se en l'estrès de la CPU. Per exemple, Linux utilitza la CPU com a mètrica per a determinar la implicació de la CPU. En canvi, hi ha altres possibilitats com ara utilitzar el nombre de milions d'instruccions executades per segon (MIPS). La càrrega de treball es pot descompondre en dues parts: la càrrega de treball del xip (el rendiment depèn de la freqüència de la CPU) i la càrrega de treball de fora del xip (el rendiment no depèn de la freqüència de la CPU).

L'impacte del canvi de freqüència de la CPU té una certa repercussió en el temps d'execució, que es pot modelitzar de la manera següent:

$$1+\delta = \frac{T(f)}{T(f_{màx})} \approx \frac{mips(f_{màx})}{mips(f)} \approx \beta \frac{f_{màx}}{f} + (1-\beta)$$

En aquesta expressió, β quantifica el nivell d'intensitat de la càrrega de treball fora del xip, δ és la desacceleració relativa del rendiment, $T(f)$ és el temps d'execució del treball a la freqüència de CPU f, $mips(f)$ és la mitjana dels MIPS per a la freqüència de la CPU f, i $f_{màx}$ és la freqüència màxima de la CPU.

També es poden aprofitar les fases intensives de memòria per a seleccionar adequadament la freqüència de la CPU. Les fallades de memòria cau de nivell més baix són bons indicadors de si una fase d'execució és intensiva de memòria. Una execució es pot dividir en una seqüència de finestres en què cada finestra conté un nombre fix de cicles de rellotge.

Al final de cada finestra es poden determinar les característiques de la fase corresponent. Per exemple, es pot utilitzar una ràtio de fallades de memòria cau de nivell 3 (exterior) amb valor 0,4 com a llindar per a detectar fases intensives de memòria. Per sobre d'aquest llindar, la nova freqüència de la CPU f_{new} es calcula de la manera següent:

$$f_{New} = f_{Old} \times \frac{\text{Instruccions executades en la finestra activa}}{\text{Instruccions executades en l'última finestra} \times 100} \times 100$$

Si la fase no és intensiva de memòria, la CPU funciona a màxima freqüència. Hi ha altres tècniques que també combinen els MIPS amb la ràtio de fallades de memòria cau per determinar el tipus de fase.

Tal com hem vist durant la resta dels mòduls didàctics, les aplicacions en computació d'altes prestacions són aplicacions paral·leles, normalment MPI. Així doncs, moltes de les tècniques de gestió d'energia per a computació d'altes prestacions estan basades en la detecció de fases en què es pot treure profit dels mecanismes de control de potència dels diferents subsistemes, com, per exemple, la CPU.

Vegeu també

Vegeu els MPI (*message passing interface*) en el subapartat 3.2.1 del mòdul "Introducció a la computació d'altes prestacions" d'aquesta assignatura.

Una de les tècniques en aquest sentit és la de detectar fases en programes MPI en les quals s'aplicaran nivells inferiors de consum a la CPU, basats en la memòria. Aquestes fases de l'execució de l'aplicació MPI es poden reconèixer a partir de dos passos:

1) En el primer pas, el programa es divideix en bloc utilitzant dues regles:

- Qualsevol operació MPI determina el límit d'un bloc.
- Si la pressió en la memòria canvia abruptament, el límit d'un bloc succeeix en aquest canvi.

Per a implementar la primera regla, es poden interceptar crides MPI (per exemple, mitjançant la interfície PMPI), i per a la segona, es poden utilitzar les operacions per fallada de memòria cau com a indicador de la pressió en la memòria.

2) En el segon pas, els blocs s'uneixen en fases. Dos blocs adjacents s'uneixen si la pressió que exerceixen sobre la memòria està entre el mateix llindar. A partir d'això es pot aplicar un algoritme per a establir la freqüència de la CPU més apropiada que optimitzi l'eficiència energètica per aquella fase determinada.

Aquestes tècniques determinen si la CPU limita o no el rendiment del sistema a partir de les activitats de la CPU, com ara la utilització de la CPU, l'MPIS o la ràtio de fallades de memòria cau. Aquestes activitats són efectives per a les fases en què hi ha operacions externes que bloquegen el rendiment de la CPU. En canvi, en algunes situacions, el rendiment de la CPU no és el límit del rendiment de tot el sistema. Per exemple, la CPU podria estar ocupada en un bucle en espera d'altres esdeveniments i, per tant, no cal que operi a la freqüència màxima. Les tècniques anteriors no detecten aquesta situació, i es pot perdre l'oportunitat d'estalviar energia.

Anàlisi de la fase de comunicació

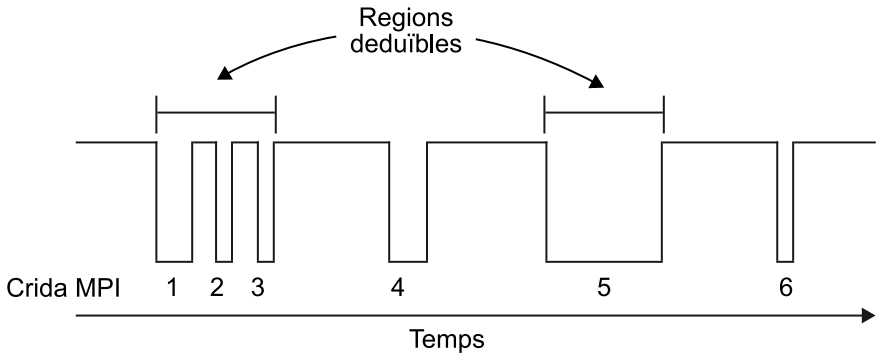
Una **execució delimitada per comunicació** succeeix quan un node n'ha d'esperar un altre per acabar la comunicació, ja sigui perquè el dispositiu és més lent o perquè s'ha donat més feina al node.

S'han desenvolupat tècniques per a reconèixer aquestes situacions i prendre mesures per a estalviar energia. A continuació, tractarem de dues de les més comunes, que, a més, ho són de la gestió d'energia per a aplicacions d'altes prestacions.

Per una banda, en l'anàlisi de regions intensives de comunicació, la fase de comunicació és un període en l'execució d'una aplicació en què la comunicació és intensiva però no és intensiva de CPU. La idea és detectar aquestes fases

en temps d'execució i aplicar DVFS a la CPU en els moments adequats per a estalviar energia. Aquesta tècnica es tracta d'interceptar i emmagatzemar les seqüències de crides MPI durant l'execució d'un programa. Un segment d'un codi d'un programa es coneix com a **regió reductible** si hi ha una concentració elevada de crides MPI (per exemple, les crides 1-3 de la figura 5) o si una crida MPI és prou llarga (per exemple, la crida 5 de la figura 5). Quan s'entra una altra vegada en una regió reductible que s'ha pogut detectar, el processador es posa a la freqüència apropiada més reduïda per a estalviar energia sense degradar el rendiment de l'execució de l'aplicació.

Figura 5. Exemple de traça d'un programa d'MPI



Les crides 1-3 formen una regió reduïble de múltiples crides juntes. La regió 5 és una regió reductible, ja que és una crida llarga. Les crides 4 i 6 no es poden reduir, ja que ni estan prou juntes ni són prou llargues.

Per altra banda, també hi ha l'anàlisi de desequilibri entre nodes. Quan la càrrega de càlcul no està ben equilibrada en tots els nodes, un node que arriba abans a un punt de sincronització ha d'esperar els altres nodes més lents. Quan succeeix això, es diu que hi ha *slack* en el node per desequilibri. Aquesta situació pot succeir diverses vegades si el codi està en un bucle. Així doncs, el que s'intenta és posar el node més ràpid a una freqüència més baixa per a conservar l'energia sense pèrdua de rendiment significativa. Es pot calcular l'*slack* d'un node com el temps d'espera en una iteració dividit per la longitud de la iteració. L'*slack* de xarxa d'un node es calcula com la diferència entre el seu propi *slack* i l'*slack* mínim dels nodes.

2.5.2. Planificació i assignació de treballs

Els principis de conservació d'energia per mitjà de la planificació i assignació de tasques desenvolupades inicialment per a clústers formats amb maquinari de gran consum també es poden aplicar a la computació d'altes prestacions. No obstant això, les solucions tècniques emprades amb èxit en els centres de dades com ara la concentració de la càrrega són ineficaces per a supercomputadors perquè les característiques que tenen de càrrega de treball i la demanda de rendiment són diferents. Ens centrarem en l'assignació de tasques entre nodes, ja que dins d'un node es poden aplicar tècniques com les vistes en relació amb el sistema operatiu.

Un exemple de gestor de sistemes de cues que té un cert suport per a la gestió d'energia en els nodes inactius és l'SLURM. Un node que roman inactiu durant un període pot ser col·locat en un estat de baixa energia i torna a l'estat normal un cop que s'assigna un nou treball en aquell node.

Per evitar un augment instantani de la demanda de potència pel fet que el nombre de nodes actius canvien ràpidament, l'SLURM pot engegar i apagar els nodes a poc a poc. No obstant això, no proporciona cap política de gestió d'energia. L'usuari decideix quan de temps ha de passar abans de canviar a mode d'energia més baixa, els estats de potència desitjats i el nombre màxim de nodes que poden canviar d'estat per minut. La gestió d'energia de l'SLURM pot ser útil, però no s'ha dut a terme gaire recerca sobre la manera d'utilitzar-la eficaçment, sobretot en la manera d'assignar tasques als nodes necessaris per a la conservació d'energia. Algunes de les iniciatives en el món de la recerca que intenten millorar les polítiques de planificació tenen en compte tècniques com ara DVFS però encara no han arribat a implantar-se en sistemes reals.

3. Computació sostenible i per a la sostenibilitat

En la *green computing* normalment es parla de desenvolupar solucions perquè la computació sigui sostenible, per exemple no superant certs límits de potència elèctrica ja que podria no estar disponible. En canvi, la computació també es pot utilitzar per a millorar la sostenibilitat del medi ambient, per exemple amb solucions per a desenvolupar xarxes elèctriques intel·ligents[21] que facin un ús i transport més útil de l'energia elèctrica, per a simular possibles emplaçaments de plaques solars o edificis en un barri, o escollir l'emplaçament de la instal·lació de manera idònia en funció del clima de la zona i els tipus de proveïdors d'electricitat locals. En aquesta segona categoria, hi podem trobar infinitat d'exemples però en aquest apartat tractarem breument del primer tipus.

(21) En anglès, *smart grid*.

Una de les primeres manifestacions del moviment de *green computing* va ser el llançament del programa Energy Star el 1992. Energy Star és una etiqueta voluntària que reconeix els productes informàtics que reïxen a reduir al mínim l'ús d'energia i augmentar al màxim l'eficiència. Energy Star s'aplica a productes com monitors d'ordinador, televisors i dispositius de control de temperatura com ara refrigeradors, aires condicionats i articles similars. Algunes de les pràctiques que s'han adoptat inclouen, per exemple, des d'apagar el monitor quan no està en ús o l'ús de monitors més eficients, fins a sistemes de refrigeració més eficients.

En aquest apartat utilitzarem dos casos d'ús per a il·lustrar diferents tipus de tècniques i dissenys de sistemes sostenibles. El primer és la tècnica anomenada *refrigeració natural*, que permet reduir dràsticament el consum elèctric del sistema, i el segon és un cas particular de microcentre de dades basat en energia solar.

3.1. Refrigeració natural

Una tendència actual en el disseny de centres de dades és la utilització del mètode refrigeració natural[22] per a la refrigeració, que està tenint un gran impacte econòmic i mediambiental.

(22) En anglès, *free cooling*.

La **refrigeració natural** és un mètode econòmic que consisteix en la utilització de les baixes temperatures de l'aire exterior per a ajudar en la refrigeració de l'aigua que després és utilitzada per a refrigerar els sistemes.

Hi ha tres maneres bàsiques d'utilitzar la refrigeració natural:

1) *Strainer cycle*. La torre de refredament d'aigua es pot relacionar directament amb el flux a través del circuit de l'aigua de refrigeració. Si la torre de refrigeració està oberta, llavors es requereix un filtre per a eliminar qualsevol residu que pugui acumular dins de la torre. L'estalvi de costos s'associa a l'ús limitat de l'aigua de refrigeració, però hi ha un risc més gran de corrosió utilitzant aquest mètode.

2) *Plate and frame heat exchanger*. Amb un intercanviador de calor, es transfereix calor directament des del circuit d'aigua de refrigeració a la torre de refrigeració. L'intercanviador manté l'aigua de la torre de refrigeració separada del refrigerant que flueix a través dels serpentins de refrigeració. L'aigua de refrigeració és per tant prerefredada. L'estalvi d'energia es redueix pel refredador de càrrega i per tant hi ha una reducció en el consum d'energia. Hi ha un augment en el cost a causa de la bomba que es necessita per a compensar les diferències de pressió.

3) *Refrigeration migration*. La disposició d'una vàlvula dins l'aigua de refrigeració obre un camí directe entre el condensador i l'evaporador. El líquid del circuit de refrigeració es vaporitza i l'energia es transporta directament al condensador, on és refredat i condensat per l'aigua de la torre de refrigeració. Aquest mètode està basat en la idea que el refrigerant tendeix a desplaçar-se cap al punt més fred en un circuit de refrigeració. L'estalvi de costos associats a aquest mètode es deu a la inactivitat del compressor, ja que el ventilador i les bombes estan operatius.

Exemples d'ús de tècniques de refrigeració natural

L'empresa Google té dos centres de dades a Europa (un a Bèlgica i un altre a Finlàndia) que no utilitzen sistemes de refrigeració tradicionals sinó que es refrigeren a partir de l'aigua com a font d'energia natural i sostenible.

Un altre exemple és el SuperMUC del Centre de Supercomputació de Leibniz, que utilitza una forma nova i revolucionària de refrigeració per aigua desenvolupada per IBM. Aquest mètode està relacionat amb la tolerància en augment dels components electrònics a temperatures més elevades, que fa que es necessiti menys refrigeració i, en conseqüència, menys energia.

3.2. Cas d'ús: Parasol

Parasol és un microcentre de dades solar desenvolupat per investigadors de Rutgers, The State University of New Jersey, als Estats Units. Està format per un petit contenidor, un conjunt de panells solars i bateries. El contenidor es troba en una estructura d'acer col·locada al terrat d'un dels edificis del campus. Els panells solars es munten a la part superior de l'estructura d'acer de manera que pugui adquirir l'energia solar i protegir el contenidor del sol. El contenidor allotja dos armaris de servidors de baix consum (fins a 160) i equips de xarxa. El contenidor utilitza refrigeració natural sempre que és possible i activa l'aire condicionat convencional quan cal.

A més dels panells solars, Parasol pot obtenir energia de les seves bateries o de la xarxa elèctrica. Hi ha tres interruptors manuals que permeten configuracions diferents per al subministrament d'energia. Per exemple, es pot configurar Parasol per a operar completament fora de la xarxa elèctrica. Parasol també inclou una àmplia infraestructura de monitoratge per a poder quantificar la quantitat d'energia que s'extreu de cada font disponible i així poder implementar polítiques d'eficiència energètica. La figura següent mostra una fotografia de l'exterior de Parasol.

Figura 6. Fotografia de l'exterior del microcentre de dades Parasol a Rutgers, The State University of New Jersey

Font: pàgina web del projecte Parasol

Més informació

Podeu trobar més informació sobre el projecte Parasol en la llista de lectures recomanades de l'apartat 4.

4. Llista de lectures recomanades

En aquest apartat presentem una llista de lectures recomanades per tal de complementar aquest mòdul didàctic. Es tracta bàsicament d'articles de recerca actuals que cobreixen la majoria dels aspectes relacionats amb aquest mòdul i que ofereixen oportunitats d'aprofundir en el tema a partir de fer-ne una lectura crítica.

4.1. Conceptes generals

U. Hölzle; L. A. Barroso (2009). *The datacenter as a computer: An introduction to the design of warehouse-scale machines* (1a. ed.). Madison, Wisconsin: Morgan and Claypool Publishers.

L. A. Barroso; U. Hölzle (2007). "The case for energy-proportional computing". *IEEE Computer Society* (vol. 40, núm. 12, pàg. 33-371). Los Alamitos, Califòrnia: IEEE Computer Society Press.

T. Scogland; B. Subramaniam; W. Feng (2012, maig). "The Green500 list: escapades to exascale". *Computer science - Research and development* (pàg. 1-9). Blacksburg, Virgínia: Springer-Verlag.

4.2. Arquitectura i gestió de subsistemes

V. Pallipadi; S. B. Siddha (2007). "Processor power management features and process scheduler: Do we need to tie them together?". A: *LinuxConf Europe*.

K. Choi; R. Soma; M. Pedram (2004, 9-11 d'agost). "Dynamic voltage and frequency scaling based on workload decomposition". A: *Proceedings of International Symposium on Low Power Electronics and Design 2004* (*ISLPED '04*) (pàg. 174-179). Califòrnia, Los Angeles: Newport Beach.

I. Hur; C. Lin (2008). "A comprehensive approach to DRAM power management". A: *14th International Conference on High-Performance Computer Architecture* (*HPCA*) (pàg. 305-316).

L. Xiaodong; L. Zhenmin; Z. Yuanyuan; A. Sarita (2005, agost). "Performance directed energy management for main memory and disks". *ACM Transactions on Storage* (vol. 1, núm. 3, pàg. 346-380). Nova York: ACM.

V. Pallipadi; S. Li; A. Belay (2007). "cpuidle - Do nothing efficiently…". A: *Ottawa Linux Symposium* (*OLS '07*).

S. Siddha; V. Pallipadi; A. van de Ven (2007). "Getting maximum mileage out of tickless". A: *Ottawa Linux Symposium* (*OLS '07*) (pàg. 201-208).

V. Delaluz; M. Kandemir; N. Vijaykrishnan; A. Sivasubramaniam; M. J. Irwin (2001). "Hardware and software techniques for controlling DRAM power modes". *IEEE Transactions on Computers* (vol. 50, núm. 11, pàg. 1154-1173). Pennsilvània: Computer Science and Engineering / Pennsylvania State University.

D. Colarelli; D. Grunwald (2002). "Massive arrays of idle disks for storage archives". A: *Proceedings of the 2002 ACM/IEEE Conference on Supercomputing* (*Supercomputing '02*) (pàg. 1-11). Los Alamitos, Califòrnia: IEEE Computer Society Press.

D. Rotem; E. Otoo; S.-C. Tsao (2009). "Analysis of trade-off between power saving and response time in disk storage systems". *Fifth Work- shop on High-Performance Power-Aware Computing (HPPAC '09) with IPDPS '09*.

E. Pinheiro; R. Bianchini; C. Dubnicki (2006). "Exploiting redundancy to conserve energy in storage systems". *ACM SIGMETRICS Performance Evaluations Review* (vol. 34, núm. 1, pàg. 15-26). Nova York: ACM.

4.3. Entorns de computació d'altes prestacions

N. Kappiah; V. W. Freeh; D. K. Lowenthal (2005). "Just in time dynamic voltage scaling: exploiting inter-node slack to save energy in MPI programs". A: *ACM/IEEE Conference on Supercomputing* (*SC '05*) (pàg. 33). Raleigh: North Carolina State University.

B. Rountree, D. K. Lownenthal, B. R. de Supinski, M. Schulz, V. W. Freeh; T. Bletsch (2009). "Adagio: making DVS practical for complex HPC applications". A: *23rd International Conference on Supercomputing* (*ICS '09*) (pàg. 460-469). Nova York: ACM.

W. Freeh; F. Pan; N. Kappiah; D. K. Lowenthal; R. Springer (2005). "Exploring the energy-time tradeoff in MPI programs on a power-scalable cluster". A: *19th IEEE International Parallel and Distributed Processing Symposium* (*IPDPS '05*) (pàg. 4.1). Raleigh: North Carolina State University.

K. W. Cameron; R. Ge; X. Feng (2005). "High-performance, power-aware distributed computing for scientific applications". *Computer* (vol. 38, núm. 11, pàg. 40-47). Blacksburg, Virgínia: Virginia Polytechnic Institute and State University.

M. Y. Lim; V. W. Freeh; D. K. Lowenthal (2006). "Adaptive, transparent frequency and voltage scaling of communication phases in MPI programs". A: *ACM/IEEE Conference on Supercomputing* (*SC '06*) (pàg. 107). Nova York: ACM.

B. Rountree; D. K. Lowenthal; S. Funk; V. W. Freeh; B. R. de Supinski; M. Schulz (2007). "Bounding energy consumption in large-scale MPI programs". A: *ACM/IEEE Conference on Supercomputing* (*SC '07*) (pàg. 1-9). Athens, Geòrgia: University of Georgia.

V. W. Freeh; D. K. Lowenthal (2005). "Using multiple energy gears in MPI programs on a power-scalable cluster". A: *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (*PPoPP '05*) (pàg. 164-173). Nova York: ACM.

4.4. Centres de dades

J. Moore; J. Chase; P. Ranganathan (2005, abril). "Making scheduling «cool»: Temperature-aware workload placement in data centers". A: *Proceedings of the 2005 USENIX Annual Technical Conference* (*USENIX '05*). Berkeley, Califòrnia: USENIX Association Berkeley.

A. Verma; P. Ahuja; A. Neogi (2008). "pMapper: power and migration cost aware application placement in virtualized systems". A: *9th ACM/IFIP/USENIX International Conference on Middleware* (*Middleware '08*) (pàg. 243-264). Nova York: Springer-Verlag.

R. Nathuji; K. Schwan (2007). "VirtualPower: coordinated power management in virtualized enterprise systems". A: *ACM SIGOPS Symposium on Operating Systems Principles* (*SOSP '07*) (pàg. 265-278). Nova York: ACM.

A. Qureshi; R. Weber; H. Balakrishnan; J. Guttag; B. Maggs (2009). "Cutting the electric bill for internet-scale systems". *SIGCOMM Computer Communication Review* (vol. 39, núm. 4, pàg. 123-134). Nova York: ACM.

J. Moore; J. Chase; P. Ranganathan (2006, juny). "Weatherman: Automated, online, and predictive thermal mapping and management for data centers". A: *Proceedings of the Third IEEE International Conference on Autonomic Computing*. Washington: IEEE Computer Society Press.

X. Fan; W. Weber; L. A. Barroso (2007). "Power provisioning for a warehouse-sized computer". A: *Proceedings of the 34th Annual International Symposium on Computer Architecture (ISCA '07)*. Nova York: ACM.

4.5. Sostenibilitat

C. Li, W. Zhang, C. Cho; T. Li (2011). "SolarCore: Solar energy driven multi-core architecture power management". A: *Proceedings of the 2011 IEEE 17th International Symposium on High Performance Computer Architecture* (*HPCA '11*) (pàg. 205-216). Washington: IEEE Computer Society.

Z. Liu; M. Lin; A. Wierman; S. H. Low; L. H. Andrew (2011, desembre). "Geographical load balancing with renewables". *ACM SIGMETRICS Performance Evaluations Review* (vol. 3, núm. 39, pàg. 62-66). Nova York: ACM.

I. Goiri; W. Katsak; K. Le; T. D. Nguyen; R. Bianchini (2013, 16-20 de març). "Parasol and GreenSwitch: Managing datacenters powered by renewable energy". A: *18th International Conference on Architectural Support for Programming Languages and Operating Systems* (*ASPLOS 2013*). Houston.

N. Sharma; J. Gummeson; D. Irwin; P. Shenoy (2010, 21-25 de juny). "Cloudy computing: Leveraging weather forecasts in energy harvesting sensor systems". A: *Sensor Mesh and Ad Hoc Communications and Networks* (*SECON*). *7th Annual IEEE Communications Society Conference on* (pàg. 1-9). Boston, Massachusetts: IEEE Conference Publications.

C. Stewart; K. Shen (2009, octubre). "Some joules are more precious than others: Managing renewable energy in the datacenter". A: *Proceedings of the Workshop on Power Aware Computing and Systems*. Big Sky, Montana.

K. Ley; R. Bianchiniy; M. Martonosi; T. D. Nguyeny (2009). "Cost and energy-aware load distribution across data centers". *HotPower '09*.

Bibliografia

Barroso, L. A.; Hölzle, U. (2007). "The case for energy-proportional computing". *IEEE Computer* (vol. 40, núm. 12, pàg. 33-371).

Brown, L.; Keshavamurthy, A.; Shaohua Li, D.; Moore, R.; Pallipadi, V.; Yu, L. (2005). "ACPI in Linux. Intel open source technology center". *Proceedings of the Linux Symposium*. Ottawa.

Chari, S. (2011, juny). "IBM Blue Gene/Q: The most energy efficient solution for high performance computing, a total cost of ownership (TCO) study comparing the IBM Blue Gene/Q with traditional x86 based cluster systems including systems with graphics processing units (GPUs)". Cabot Partners.

Hölzle, U.; Barroso, L. A. (2009). *The datacenter as a computer: An introduction to the design of warehouse-scale machines* (1a. ed.). Madison, Wisconsin: Morgan and Claypool Publishers.

Intel; Microsoft; Toshiba (1996, 22 de desembre). *Advanced Configuration and Power Interface Specification*.

Kappiah, N.; Freeh, V. W.; Lowenthal, D. K. (2005). "Just in time dynamic voltage scaling: exploiting inter-node slack to save energy in MPI programs" A: *ACM/IEEE Conference on Supercomputing* (*SC '05*) (pàg. 33). Raleigh: North Carolina State University.

Lim, M. Y.; Freeh, V. W.; Lowenthal, D. K. (2006). "Adaptive, transparent frequency and voltage scaling of communication phases in MPI programs". A: *ACM/IEEE Conference on Supercomputing* (*SC '06*) (pàg. 107). Nova York: ACM.

Pallipadi, V.; Li, S.; Belay, A. (2007). "cpuidle - Do nothing efficiently…". A: *Ottawa Linux Symposium* (*OLS '07*).

Rountree, B.; Lowenthal, D. K.; Funk, S.; Freeh, V. W.; Supinski, B. R. de; Schulz, M. (2007). "Bounding energy consumption in large-scale MPI programs". A: *ACM/IEEE Conference on Supercomputing* (*SC'07*) (pàg. 1-9). Athens, Geòrgia: University of Georgia.

Scogland, T.; Subramaniam, B.; Feng, W. (2012, maig). "The Green500 list: escapades to exascale" (pàg. 1-9). *Computer science - Research and development*. Blacksburg, Virgínia: Springer-Verlag.