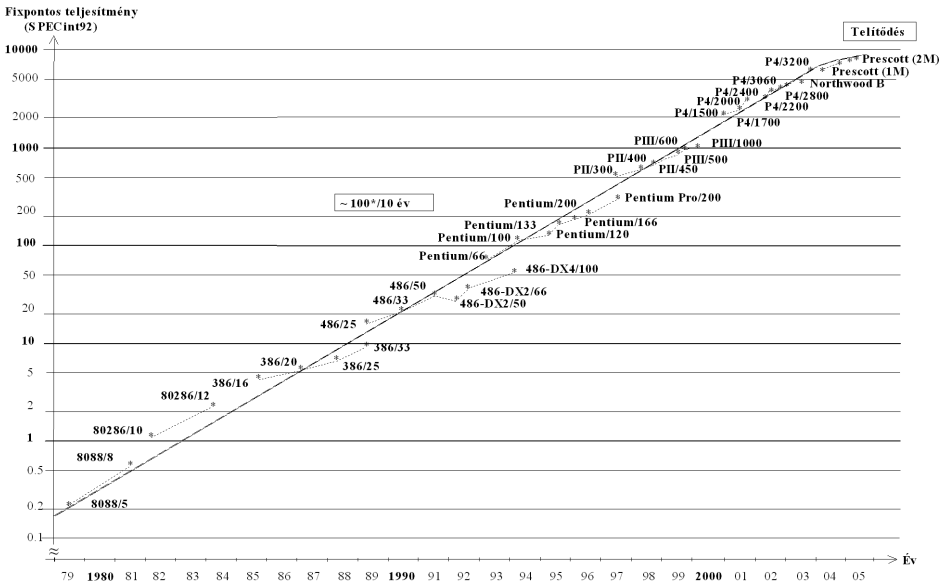
Alkalmazott Matematikai Lapok **24** (2007), 265-275.

KORSZAKVÁLTÁS A PROCESSZOROK FEJLŐDÉSÉBEN

SIMA DEZSŐ

1. Bevezetés

Több mint két évtizede – közel napjainkig – a processzorok teljesítménye megdöbbentően gyors ütemben folyamatosan emelkedett, amint ezt a világpiacon vezető piaci részesedésű Intel x86 család fixpontos teljesítményének változását bemutató 1. ábra illusztrálja. E szerint több mint két évtizeden át az Intel x86 processzorok fixpontos teljesítménye 10 évente mintegy megszázszorozódott. Ugyanakkor az elmúlt néhány évben jól kivehetővé vált egy új fejlődési szakasz, melyben a processzor teljesítmények növekedési üteme lényegesen lelassult és a növekedési görbe mindinkább egy telítődési görbéhez kezdett hasonlítani. De mi is játszódott le az elmúlt években, milyen okok idézték elő ezt a gyökeres változást?



1. ábra. Az Intel x86 család fixpontos teljesítményének növekedése [1], [2]

Az okok meghatározásához jó kiindulásul szolgál a processzorok abszolút műveleti teljesítményének vizsgálata.

Utasítás szinten a processzorok abszolút műveleti teljesítménye, azaz az időegység alatt végrehajtott műveletek átlagos száma (P_O) az alábbi összefüggéssel írható le [3]:

$$P_O = f_c \cdot IPC \cdot OPI \cdot \eta,$$

ahol f_c órafrekvencia, IPC a ciklusonként kibocsátott utasítások átlagos száma, OPI utasításonként a műveletek átlagos száma, és η a spekulatív végrehajtás hatékonysága, azaz az eredményesen végrehajtott utasítások száma/kibocsátott utasítások száma.

A fenti összefüggés két komponens szorzataként is felírható:

$$P_O = f_c \cdot E_p,$$

ahol $E_p = IPC \cdot OPI \cdot \eta$ az egy óraciklus alatt eredményesen végrehajtott műveletek átlagos számát, azaz a processzor hatékonyságát tükrözi.

E szerint utasítás szinten a processzorok teljesítménye alapvetően két úton fokozható; vagy a processzor órafrekvenciájának (f_c) vagy a hatékonyságának (E_p) a növelésével. A következőkben vizsgáljuk meg, hogy e teljesítmény összetevők növekedési üteme idővel hogyan változott az Intel processzorok példáján!

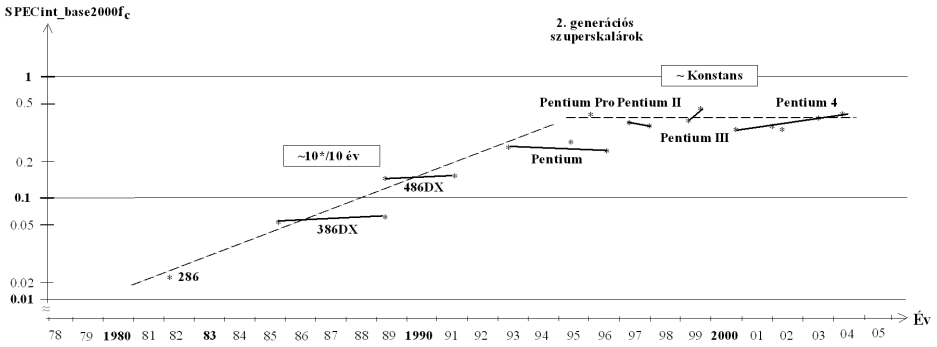
2. Az első hatékonysági korlát – az általános alkalmazásokban utasításszinten rendelkezésre álló párhuzamosság kimerülése

Elsőként tekintsük a processzorok hatékonyságát, és összpontosítsuk vizsgálódásunkat az általános célú (fixpontos) alkalmazásokra. Ez esetben a processzorok hatékonysága az általános célú fixpontos mérőprogramok (pl. SPECint92, SPECint95, SPECint_base2000) publikált eredményeinek [1] azonos órafrekvenciára vonatkoztatott értékeivel jellemezhető. Tekintettel arra, hogy az egyes mérőprogramok által szolgáltatott eredmények egymástól jó közelítéssel csak egy konstansban különböznek, nincs jelentősége annak, hogy mely fixpontos mérőprogramot vesszük alapul. A processzorok hatékonyságának vizsgálatához válasszuk a SPECint_base2000 mérőprogramot, és normáljuk a publikált teljesítményértékeket 1 MHz órafrekvenciára. Ekkor valamely processzor hatékonysága (E_p) a publikált SPECint_base eredményekből az alábbiak szerint határozható meg:

$$E_p = \text{SPECint\_base2000}/f_c \qquad [1/\text{MHz}].$$

A processzorok általános célú programok futtatása esetén mért hatékonyságában idővel bekövetkezett változásokat jól szemlélteti az Intel x86 család egymást követő processzorainak hatékonyságát feltüntető 2. ábra.

A 2. ábra szerint az x86 család processzorainak hatékonysága időben jó közelítéssel két markánsan eltérő szakaszra bontható; az első időszakban, azaz a 2. generációs szuperskalár Pentium Pro megjelenéséig a processzorok hatékonysága jelentős mértékben, tízévente közel egy nagyságrenddel nőtt, míg az azt követő



2. ábra. Az Intel x86 processzorok hatékonyságának időbeli változása fixpontos alkalmazások esetén [1], [2]

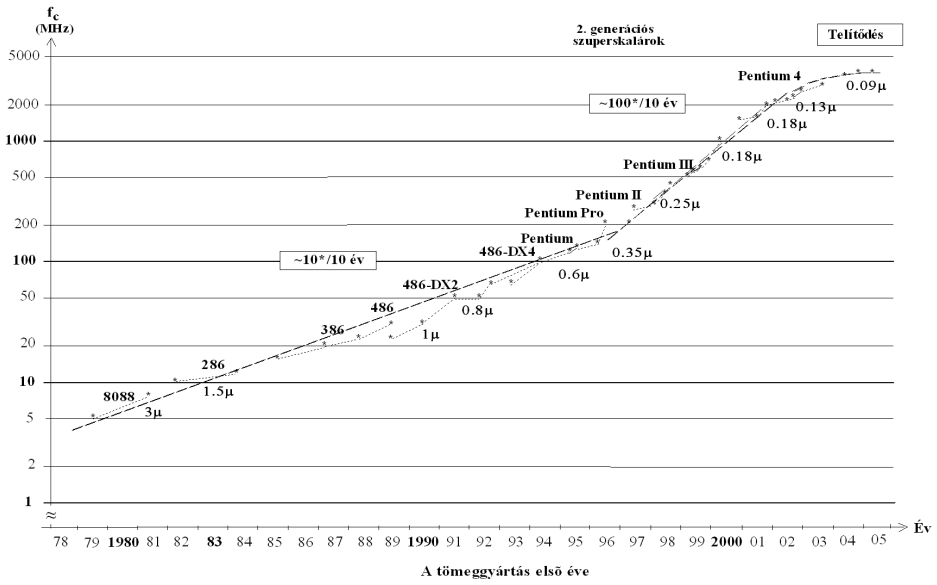
időszakban az L2 gyorsítótár processzorlapkára való integrálása (Pentium III) és kapacitásának jelentős növelése (Pentium 4) ellenére közelítőleg állandó maradt. Az első szakaszban a hatékonyság emelésének két forrása volt, egyrészt a szóhossz növelése 16 bitről (I 286) 32 bitre (I 386), másrészt a feldolgozás párhuzamosságának a fokozása, azaz az egy óraciklus alatt eredményesen feldolgozott utasítások számának (IPC) a növelése, négy lépésben. Az első lépés az időbeli párhuzamosság, azaz a futószalag-alapú feldolgozás bevezetése volt (I 386), a másodikat a futószalag technika keltette feldolgozási szűk keresztmetszetek feloldása jelentette az elágazásbecslés és a gyorsítótárak alkalmazása révén, majd az időbeli párhuzamosságban rejlő lehetőségek kimerülését követően a következő lépés szükségszerűen a kibocsátási párhuzamosság, azaz a szuperskalár feldolgozás bevezetése volt. A hatékonyság növelésének utolsó momentumát a szuperskalár kibocsátás bevezetésével adódó feldolgozási szűk keresztmetszetek feloldása képezte megfelelő technikák alkalmazásával, mint pl. az elődekódolás, utasításvárakoztatás, regiszter átnevezés, több portos, nem blokkolódó gyorsítótárak stb. [3]. Az irodalomban megjelent áttekintések szerint a második generációs szuperskalárok megjelenéséig hasonló ütemben növekedett más processzorcsaládok (Alpha, MIPS, PA-RISC, POWER, PowerPC, SPARC) egymást követő modelljeinek fixpontos teljesítménye ill. hatékonysága is [4].

A második generációs szuperskalárok megjelenésével viszont új korszak köszöntött be a processzorok fejlődésében, mivel a második generációs 3-utasítás/óraciklus feldolgozási szélességű CISC magok (mint pl. a Pentium Pro) vagy a 4-utasítás/óraciklus szélességű RISC magok érdemben már kiaknázzák az általános célú programokban utasítás szinten rendelkezésre álló 4-8 utasítás/ciklusnyi párhuzamosságot [5]. Következésképpen a második generációs szuperskalárokat követően a processzorok hatékonysága általános célú alkalmazásokban a feldolgozási szélesség növelésével már nem volt lényegesen tovább fokozható, azaz a második generációs szuperskalárokkal kezdődően a processzorok teljesítményének további növelése általános célú alkalmazásokban egy hatékonysági korlátba ütközött.

Itt megjegyezzük, hogy dedikált alkalmazásokban, pl. szerver környezetekben vagy a 90-es évek második felétől kezdve rohamosan elterjedő multimédiás, ill. 3D-s alkalmazásokban utasításszinten is még további lehetőségek nyíltak a processzorok hatékonyságának a fokozására. Szerver környezetekben utasítás szinten lényegesen nagyobb mérvű funkcionális párhuzamosság állhat rendelkezésre, míg a multimédiás és grafikus alkalmazások utasításonként több művelet (OPI) végrehajtását is lehetővé adatpárhuzamosságot kínálnak. Ez utóbbi lehetőségeket hasznosítják az elmúlt évtized végén megjelent fix- és lebegőpontos SIMD utasításokkal kiegészített harmadik generációs szuperskalárok (pl. a Pentium III, Pentium 4, Athlon, Power 3).

3. Az órafrekvenciák rohamos növelése, a következmények

A következőkben vizsgáljuk meg, hogy a tekintett időszakban a processzor teljesítmények növelésének másik lehetséges dimenziójában; az órafrekvenciák növelésében, milyen változások következtek be az Intel x86 processzorok példáján (ld. a 3. ábrát). Itt megjegyezzük, hogy az Intel x86 processzorok esetében az alábbiakban megfogalmazott megállapítások a processzorok fejlődésére vonatkozóan általában is érvényesek.



3. ábra. Az Intel x86 processzorok órafrekvenciájának növekedése [2]

A 3. ábrában az órafrekvenciák növekedési üteme három jól elkülönülő szakaszra tagolódik. Az első szakaszban, a 2. generációs szuperskalár Pentium Pro

megjelenéséig, a processzorok órafrekvenciája közelítőleg egy nagyságrend/10 év ütemben nőtt. Ebben a szakaszban a processzorok hatékonyságának és órafrekvenciájának növelése közel azonos mértékben (10*/10 év) járult hozzá a processzor teljesítmények közelítőleg 100*/10 év ütemű fokozásához. Ezt követően azonban, az általános alkalmazásokban utasításszinten rendelkezésre álló párhuzamosság egyre inkább kimerült, és így a processzorok hatékonyságának további növelése korlátokba ütközött és megállt. Ezért a 2. generációs Pentium Pro-t, ill. általában a 2. generációs szuperskalárokat követően a processzor teljesítmények növelésének alapvető forrása az órafrekvencia emelése lett, és a processzorok fejlődésében egy új korszak sejlett fel – az órafrekvenciák rohamos növelésének időszaka.

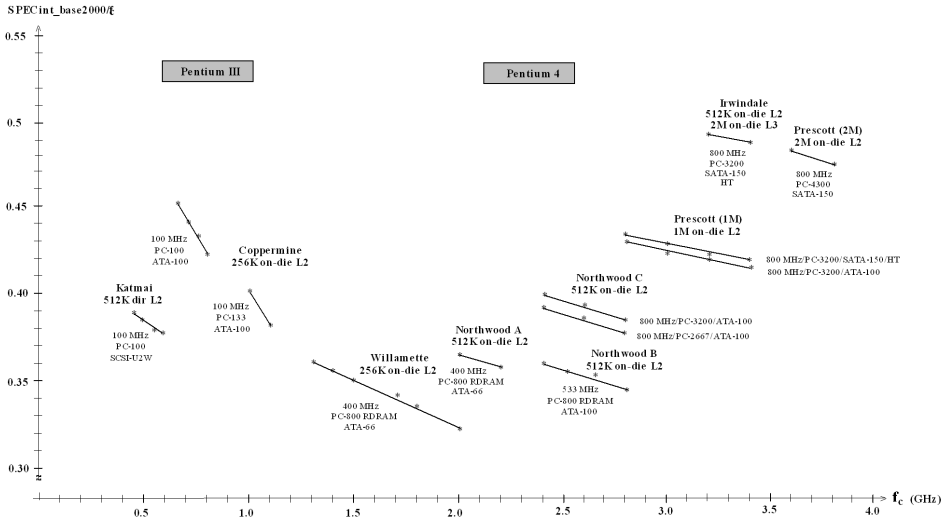
A processzorok órafrekvenciájának növelésére alapvetően két lehetőség kínálkozik; a gyártási technológia fejlesztése az áramköri struktúrák méretcsökkentése érdekében (jellemzően 2-évente 0,7x-es mértékben), valamint a futószalag fokozatok kritikus logikai hosszának csökkentése a futószalag fokozatok számának emelésével, azaz a futószalag hosszának a növelésével. A processzorgyártó cégek természetesen mindkét lehetőséggel egyidejűleg éltek, így például az Intel közel két éves ritmusban vezette be egymást követően a 0,25, 0,18, 0,13 majd a 0,09 μm-es gyártási technológiákat, és ezzel párhuzamosan a futószalag fokozatok kritikus logikai hosszának csökkentése érdekében egymást követő processzoraiban jelentősen növelte a futószalagok hosszát. Amíg a Pentium Pro alap-futószalagja mindössze 12 fokozatból állt, az órafrekvenciák erőteljes növelésének szándékával kifejlesztett Netburst architektúrát megvalósító Pentium 4 Willamette és Prescott magok alap-futószalagja már megközelítőleg 20, ill. 30 fokozatú lett. A technológiai és mikroarchitektúrális fejlesztések együttes eredményeképpen Intel a 2. generációs Pentium Pro-t követően processzorai órafrekvenciáját drasztikusan, tíz évre vonatkoztatva közel 100-szoros mértékben tudta növelni (ld. a 3. Ábrát).

Az órafrekvenciák erőltetett ütemű, erőteljes növelése az elmúlt évtized második felében három sarkalatos fejlődési korlát kiváltó okává vált, nevezetesen a második hatékonysági-, a disszipációs- és a párhuzamos buszok frekvencia korlátjának megjelenéséhez vezetett. E kérdéseket részletezzük a következő fejezetekben.

4. A második hatékonysági korlát – a processzor és a processzort kiszolgáló alrendszerek közötti sebességolló kinyílása

A 2. generációs szuperskalárokat követően az órafrekvenciák rohamos, tízévente mintegy 100-szoros mértékű növekedésének időszakában a mikroarchitektúra egyes kiszolgáló alrendszereinek (operatív tár, gyorsítótárak, processzorbusz) a „sebességnövekedése" egyre kevésbé tudta követni a processzorok igen gyors sebességnövekedését, és így egy egyre táguló sebességolló nyílt ki a processzor és egyes kiszolgáló alrendszerei között. A legeklatánsabb sebesség különbség az operatív tár tekintetében alakult ki, mely egyrészt az operatív tár ciklusokban mért elérési idejének folyamatos növekedésében, másrészt a memóriák átviteli rátájának a processzorok

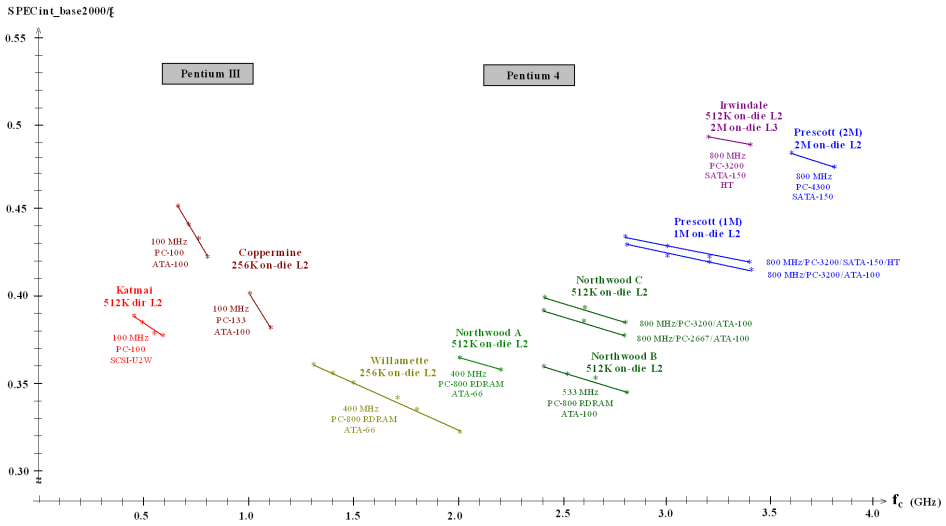
órafrekvenciájához képest lényegesen lassabb növekedési ütemében nyilvánult meg. E cikkben terjedelmi korlátok miatt nem tárgyaljuk kvantitatívan e kérdéseket és mindössze arra szorítkozunk, hogy bemutassuk a fent említett, egyre táguló sebességolló kihatását a processzorok hatékonyságára, a Pentium III és a Pentium 4 processzorcsaládok példáján (ld. a 4. ábrát).



4. ábra. Az Intel Pentium III és a Pentium 4 processzorcsalád egyes modelljeinek hatékonysága [1], [2]

Az ábra szerint növekvő órafrekvenciákon, a GHz tartományban, a kiszolgáló alrendszerek egyre nagyobb mérvű relatív elmaradása miatt a mikroarchitektúra hatékonysága azonos processzor paraméterek mellett számottevően csökken. Például miközben a Pentium III Coppermine magok órafrekvenciája $0,65$ GHz-ről $1,1$ GHz-re nőtt, hatékonyságuk a kezdeti mintegy $0,45$-ös értékről közelítőleg a $0,38$-as értékre esett vissza, annak ellenére, hogy a magasabb frekvenciájú magok esetén a processzorbusz sebességét 100 MHz-ről 133 MHz-re növelték. Az ábra egyúttal azt is érzékletesen szemlélteti, hogy a mikroarchitektúra hatékonyságának romlása elsődlegesen az L2 gyorsítótár méretének a növelésével másodlagosan a processzorbusz vagy a memória átviteli rátájának az emelésével részlegesen kompenzálható. De mindez nem változtat azon az alapvető tényen, hogy GHz tartományban az órafrekvencia növelése egyre nagyobb mértékű hatékonyságcsökkenést okoz, és ennek következtében az órafrekvencia növelésével csupán egyre csökkenő mértékű teljesítménytöbblet érhető el.

Itt megjegyezzük, hogy a 2. generációs szuperskalárokat követően, az órafrekvenciák erőteljes növelésének időszakában az eleve nagyobb órafrekvenciákon működő, de kisebb hatékonyságú RISC processzorok szükségszerűen nagyobb hatékonyságveszteséget szenvedtek el, mind az alacsonyabb órafrekvenciákon működő



5. ábra.

de hatékonyabb CISC processzorok. Következésképpen míg a processzorteljesítmények versenyében általános alkalmazások esetén az elmúlt évtized közepén a RISC processzorok (elsődlegesen az Alpha család) voltak az élen, az évtized második felében a RISC processzorok a teljesítmény versenyben egyre inkább háttérbe szorultak a CISC processzorokkal szemben [6], és a teljesítmény versenyben a CISC processzorok kerültek az első helyre. Döntően emiatt az elmúlt évtized végén a legtöbb RISC gyártó beszüntette RISC családja (Alpha, MIPS, PA-RISC, PowerPC) továbbfejlesztését, és mindössze két RISC család maradt versenyben: IBM POWER, ill. SUN UltraSPARC családja.

5. A hőtermelési korlát

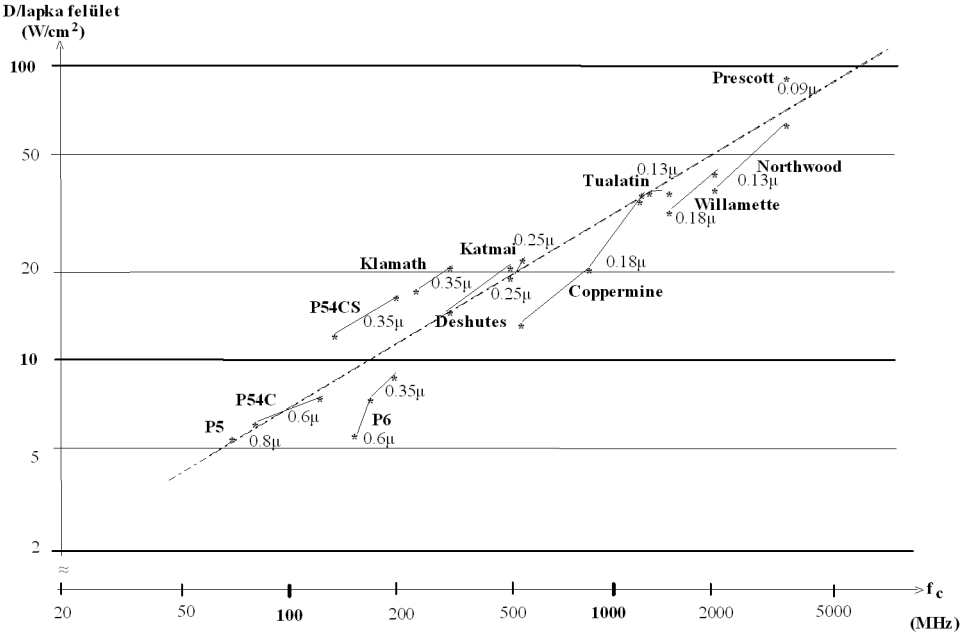
A processzorok hőtermelésének mértéke az (1) összefüggéssel adható meg [7]. Eszerint a hőtermelés dinamikus és statikus komponensekből áll, ahol a dinamikus komponens a kapuk összesített elosztott kapacitásának feltöltéséből és kisütéséből adódik, míg a statikus komponens a szivárgási áramok miatti hőtermelést jeleníti meg.

$$D = A \cdot C \cdot V^2 \cdot f_c + V \cdot I_{\text{leak}} \tag{1}$$

ahol A az aktív kapuk részaránya, C a kapuk összesített elosztott kapacitása, V tápfeszültség, f_c órafrekvencia, I_{leak} szivárgási áram.

A fenti összefüggés szerint az órafrekvenciák növekedésével a processzorok dinamikus hőtermelése lineárisan nő (egyébként azonos paraméterek mellett). A di-

namikus hőtermelés kiegészül a szivárgási áramok miatt megjelenő statikus hőtermeléssel. Az Intel x86 processzorok eredő relatív hőtermelését az órafrekvenciák függvényében az 6. ábra szemlélteti.



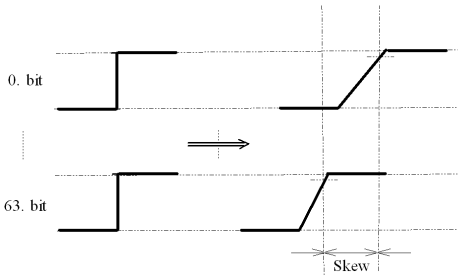
6. ábra. Intel x86 processzorok fajlagos hőtermelése [2]

E szerint magasabb órafrekvenciákon a hőtermelés is rohamosan nő és a $2,8-3,4$ GHz órafrekvenciákon bejelentett Prescott magok fajlagos, 1 cm^2 lapkafelületre eső hőtermelése már megközelíti a 100 Watt/cm^2 értéket. Léghűtést feltételezve a 100 Watt/cm^2 körüli fajlagos hőtermelés viszont már nehezen kezelhető hűtési problémákat okoz, így a Pentium 4 órafrekvenciájának növelése egyre keményebb határokba ütközött. E miatt megtorpant a Pentium 4 család korábban imponálóan gyors órafrekvencia növekedése (ld. a 3. ábrát) és Intel a korábban már bejelentett 4 GHz, ill. nagyobb órafrekvenciájú Pentium 4 modellek visszavonására, sőt a Netburst architektúra továbbfejlesztésének leállítására [7], valamint tervezési filozófiájának gyökeres módosítására kényszerült [9].

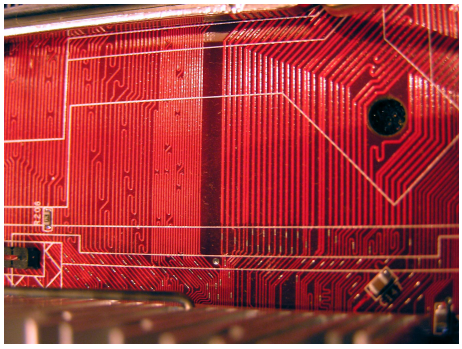
6. Párhuzamos buszok frekvenciakorlátja

Az órafrekvenciák GHz tartományban történő növelésekor egy párhuzamos busz egyes bitvezetékei között már érzékelhető futási idő különbségek (skew) jelentkeznek a bitvezetékek hosszbeli eltérése, az egyes bitvezetékek eltérő kapacitív

jellemzői által okozott jelmeredekség eltérések és a jelenlévő zajok miatt; melyek növekvő frekvenciákon – az impulzus szélességhez viszonyítva egyre dominánsabbá válnak (ld. a 7. ábrát).

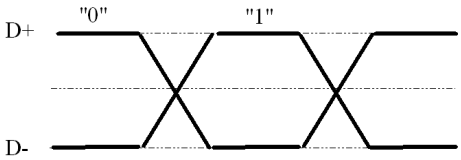


7. ábra. Párhuzamos buszok bitvezetékei közötti futási idő különbségek (skew)



8. ábra. A processzor bitvezetékei között fellépő futási idő különbségek kiegyenlítése az MSI 915G Combo alaplapon

Annak ellenére, hogy az alaplapok tervezésénél nagy figyelmet fordítanak a processzorbusz egyes bitvezetékei közötti futási idő különbségek minél jobb kiegyenlítésére (ld. a 8. ábrát), GHz tartományban a processzorbuszok sebességének a növelése az előzőekben vázolt problémák miatt egyre nagyobb nehézségekbe ütközik, és egyre nyilvánvalóbban megjelenik a párhuzamos buszok sebességkorlátja. Ezért a párhuzamos processzor buszokat napjainkban szükségszerűen egyre inkább felváltják a nagy sebességű soros buszok (pl. a Hypertransport busz). A soros buszok bitenként két vezetéket használva differenciális, kis amplitúdó váltású (néhány száz mV) jelátvitelt valósítanak meg (ld. a 9. ábrát).



9. ábra. Jelátvitel soros buszon

A gyors (egy vagy néhány Gbit/sec sebességű), egyszerűen skálázható soros buszok rohamos elterjedésének másik oka a periféria buszok vonatkozásában (USB, PCI-Express, SATA, SAS stb.) a vezetékszám drasztikus redukciójával elérhető ráfordítás csökkentés.

7. Összegzés

Általános célú alkalmazásokban az utasításszinten rendelkezésre álló párhuzamosság a második generációs szuperskalárok megjelenésével már a 90-es évek derekán kimerült. Az ezt követő közel egy évtizedben a processzor teljesítmények fokozásának színtere az órafrekvenciák intenzív növelése lett, de az elmúlt néhány évben bizonyossá vált, hogy ez az út az órafrekvenciák növekedésével egyre világosabban kirajzolódó három korlát miatt tovább már nem járható. A hatékonyság növelését célzó hardver többletráfordítások egyre csökkenő mértékben térülnek meg, a fokozódó hőtermelés mértéke túllépi a léghűtéssel kezelhető tartományt, a párhuzamos buszok egyes bitvezetékei között fellépő futási idő különbségek (skew) egyre inkább megközelítik a ciklusidőt, és így bekorlátozzák a buszfrekvenciát. Más megfogalmazásban: a processzorteljesítményeknek az órafrekvenciák intenzív növelésére alapozott stratégiája az elmúlt években hatékonysági, hőtermelési és buszfrekvencia növelési korlátokba ütközött.

E változások hatására a fejlesztések színtere az utasítás szintről a szálszintre tevődött át, az órafrekvenciák növelése érdekében alkalmazott hosszú futószalagokra (20-30 fokozat) alapozó processzorokat szükségszerűen felváltják a közepes (10-15) fokozatszámú, alacsonyabb órafrekvenciájú, lassabb, de hatékonyabb mikroarchitektúrájú többmagos processzorok, míg a párhuzamos processzorbuszokat a sebességkorlátok elérése miatt, ill. a párhuzamos periféria buszokat a ráfordítás csökkentése érdekében kiszorítják a gyors, egyszerűen skálázható soros buszok.

Hivatkozások

[1] SPEC CPU92, CPU95, CPU2000 results,

http://www.spec.org

[2] *Microprocessor Quick Reference Guide*,

http://www.intel.com/pressroom/kits/quickref.htm

[3] D. Sima: *„Decisive Aspects in the Evolution of Microprocessors"*, *Proceedings of IEEE*, Vol. 92. **No. 12**, pp. 1896–1923, (December 2004)

[4] J. Birnbaum: *„Architecture at HP: Two Decades of Innovation"*, *Microprocessor Forum*, San Jose, California, (October 14, 1997),

http://www.hpl.hp.com/speeches/mpforum.html

[5] D.W. Wall: *„Limits of Instruction Level Parallelism"*, *Proc.* 4^{th} *Int. Conf., Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pp. 176–188, (1991)

[6] *X86-64 Technology White Paper*, AMD Inc., Sunnyvale, CA, (2000),

http://www.amd.com/us-/assets/content_type/white_papers_and_tech_docs/x86-64_wp.pdf1

[7] N. S. Kim et al., *„Leakage Current: Moore's Law Meets Static Power"*, *Computer*, Vol. 36, **No. 12**, pp. 68–75, (December 2003)

[8] E. Smith: *Intel kills off 4 GHz project, moves team to multicore*,
http://www.geek.com/news/2004Oct/bch20041015027427.htm

[9] R. Ronen: *„The Thermal Wall: where it came from and how to live with it?"*, *10th Intel EMEA Academic Forum*, (May 2005),
http://download.intel.com/corporate/education/EMEA/academicforum/keynotes/Ronen

(Beérkezett: 2006. június 3.)

SIMA DEZSŐ
BUDAPESTI MŰSZAKI FŐISKOLA
NEUMANN JÁNOS INFORMATIKAI KAR
1034 BUDAPEST, BÉCSI ÚT 96/B
Email: sima@bmf.hu

THE DAWN OF A NEW ERA IN PROCESSOR EVOLUTION

Dezső Sima

The birth of second generation superscalars heralded a new age in processor evolution, since these wide superscalars already utilized most of the instruction level parallelism available in general purpose applications. As a consequence, the previous, approx. 10-fold-per-decade increase of processor efficiency leveled off. Along the main road of evolution designers addressed this crucial challenge by aggressively raising clock frequencies by a nearly 100-fold-per decade rate in order to maintain an overall approx. 100-fold-per decade performance increase. However, such an aggressive boosting of clock frequencies inevitably triggered intricate design problems in the GHz range, leading to three basic limitations in increasing performance: core efficiency, dissipation and skew walls, all contributing to the leveling off in core frequencies witnessed during the last few years. On the other hand, however, available complexity could be raised further exponentially, in accordance with Moore's law, which paved the way to a new era of processor evolution, marked by recent power-aware multicore and multithreaded designs. Our paper focuses on the three performance walls mentioned above.