Kiszolgálók teljesítményének növelése

Hogyan növeld kiszolgálód teljesítményét intelligens hálózati kártyára Linux-alapú kiszolgálóról a TCP/IP-verem áthárításával.

Növelni szeretnéd felsőosztályú Linux-kiszolgálód teljesítményét? Rendszered nagy sebességű hálózathoz kapcsolódik? Kiszolgálóidnak túl sok erőforrását emészti fel a TCP/IP-verem és az ethernetkeretek kezelése? Ezek napjaink legáltalánosabb hálózattal összefüggő gondjai, melyet a TCP/IP-forgalomnak az elmúlt évtizedben történt drámai növekedése okoz mind az Interneten, mind a nagyvállalati hálózatokon, és ez a forgalom nagy valószínűséggel csak emelkedni fog. Az Internet világméretű növekedése és az új hálózati adattároló módszerek (például iSCSI) következtében a forgalom és a sebesség még inkább nőni fog. Bár a processzorok elképesztő iramban fejlődnek, a hálózat fejlődése még ezt is túlszárnyalja, rákényszerítve a processzorokat, hogy erőforrásaikat elsődleges feladataik helyett a hálózati csomagok kezelésére pazarolják. Az Intel egy olyan intelligens hálózati kártya (network interface card – NIC) prototípust fejlesztett ki, amellyel a TCP/IP-verem a Linux-kiszolgálóról teljes egészében erre az iNIC-re irányítható.

Az iNIC felépítése

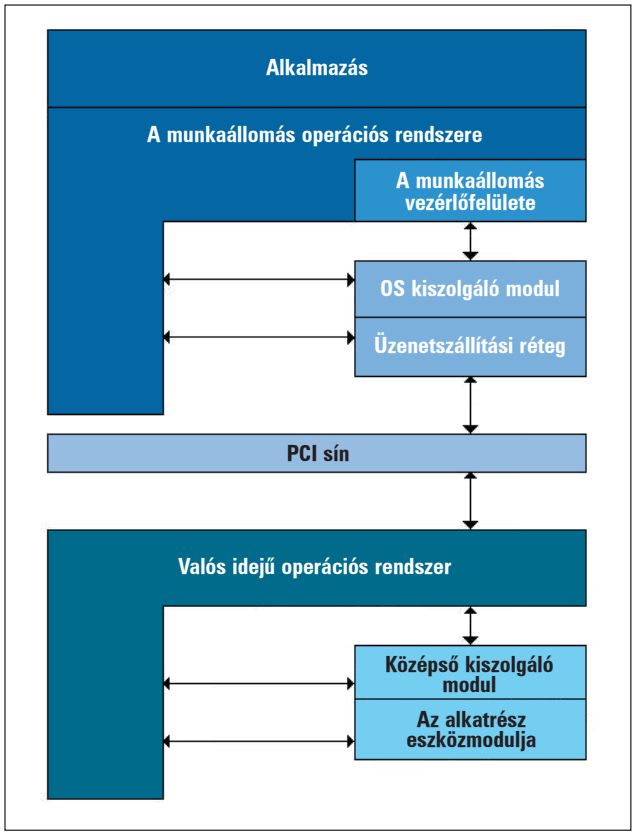
A hálózati kártyán valós idejű operációs rendszer (Real-Time OS – RTOS) fut teljes 4-es változatú TCP/IP-támogatással. Az iNIC-en a hálózati csomagokat egy I/O processzor (IOP) dolgozza fel a gazdaprocesszor tehermentesítését biztosítva. A felosztás elvégzéséhez a gazdagép részéről a legcsekélyebb logika szükségeltetik, hogy a hálózati csomagokat átirányítsa az iNIC-re.

Foglalatok áthárítása

Ez a módszer az intelligens I/O (I2O) szerkezetre épül, amelyet a Linux 2.4-es változata már támogat. Az *1. ábra* az I2O alapfelépítését tartalmazza. Az I2O-megvalósítás üzenetközvetítő módszer a gazda operációs rendszer (OS) és az IOP-on található I/O-eszközök között. Az IOP-on egy intelligens RTOS (IRTOS) fut, mely a kapcsolódó eszközök eszközmeghajtó moduljait (DDM-ek) tartalmazza. A hordozhatóság végett az I2O osztott eszközmeghajtó modellt használ. A megvalósítás minden eszközcsoporthoz alapvető üzenetrendszert határoz meg (pl: helyi hálózat, szalag, lemez). A gazda operációs rendszer is egy előre meghatározott üzenetrendszer alapján tartja a kapcsolatot az IOP-on található eszközmeghajtókkal. Ezeket az I2O-üzeneteket az eszközmeghajtó fordítja le alkatrészjellemző parancsokra. A gazdagépnek egy I2O-eszközzel való kapcsolattartásához eszközmeghajtóval kell rendelkeznie, amely a gazda operációs rendszer eszközparancsait I2O osztályszintű parancsokra fordítja le. Ezt a modult a gazdagépen operációsrendszer-modulnak (OSM) nevezik.

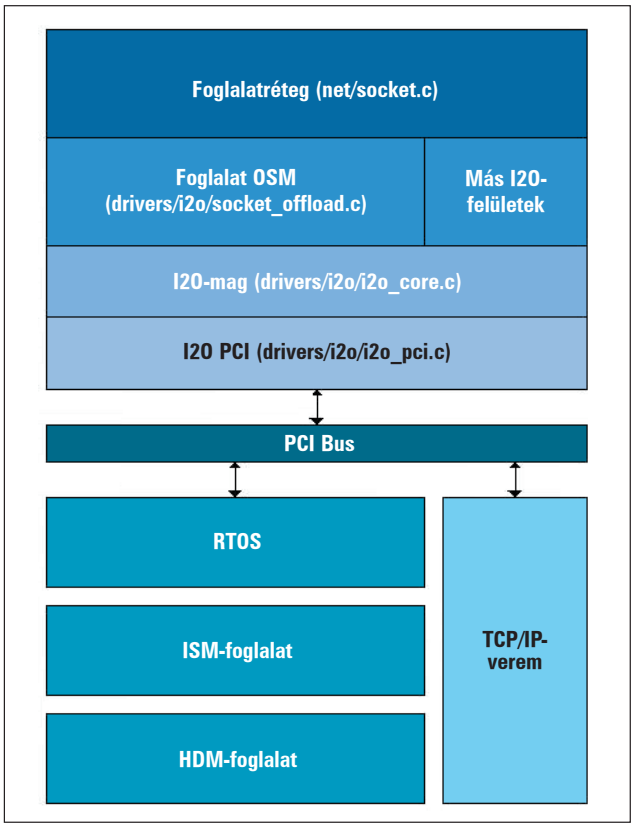
A megoldás az architektúra egy kiterjesztésére épül egy foglalati (socket) osztály létrehozásával. A teljesítménynövelés és a késés elkerülése érdekében – melyet többnyire az osztott eszközmeghajtó modell okoz – az I2O felépítésén változásokat hajtottak végre.

A foglalati osztály határozza meg a gazda operációs rendszer és az eszközmeghajtó kapcsolattartásához szükséges üzeneteket. A két meghajtó, az OSM és a DDM, azaz az operációsrendszer-modul és az eszközmeghajtó modul kétszintű üzenetközvetítő csatornán keresztül tartja a kapcsolatot. Az üzeneti réteg indítja el a kapcsolattartási folyamatot, a szállítási réteg pedig meghatározza, hogy az üzenetek hová menjenek. A DDM két almodulból épül fel: a középfokú szolgáltatás modulból (ISM) és a

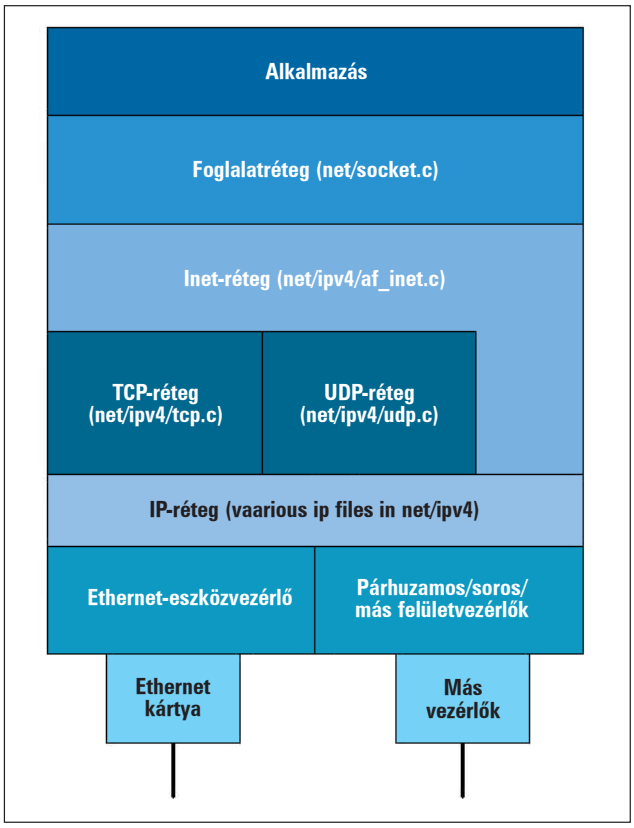


1. ábra I2O-szerkezet

gépi eszközmodulból (HDM). A 4-es változatú TCP/IP-vermet teljes egészében az ISM szolgáltatja. A HDM az iNIC eszközmeghajtója. A foglalati OSM semmilyen más Linux-eszközmeghajtóhoz nem hasonlítható. A normál hálózati kártyák meghajtói protokollfüggetlenek és a Linux-rendszermaggal a hálózati alkalmazások programozói felületén keresztül tartják a kapcsolatot. A foglalati OSM ezzel ellentétben közvetlenül a programozói felület alatti szinttel tartja a kapcsolatot – ezáltal válik lehetővé, hogy a szükséges foglalati szolgáltatások az IOP-hoz kerüljenek, amelyen a foglalati szintű ISM fut. A foglalati OSM lecseréli TCP/IP-verem által a rendszermagnak nyújtott szolgáltatásokat, így biztosítva a Linux-rendszermaghoz a szükséges felületeket, valamint a foglalati kérelmeket; az adatokat az iNIC-en futó TCP/IP-veremnek továbbítja.



2. ábra Foglalati áthárító szerkezet



3. ábra A Linux TCP/IP-verme

A foglalati OSM

Az OSM a következő alrendszerekre van felosztva: felhasználói felület, üzenetközvetítő felület, rendszermagfelület és memóriakezelő alrendszer.

A felhasználói felület lecseréli a rendszermag `af_inet` foglalati rétegét. Ez az új réteg pontosan olyan felületet biztosít, mint az eredeti. Az üzenetközvetítő réteg felel a foglalati áthárító rendszer működéséért, biztosítja az elindítását és az irányítását, és a felhasználói foglalati kéréseket foglalati üzenetekre fordítja le.

Az OSM számára a rendszermag felülete biztosít rendszermag-szolgáltatásokat. Ez az az OSM-csatlakozási pont, ahonnan az OSM TCP/IP-veremszolgáltatásokat biztosít a rendszermagnak. Ezt az alrendszert úgy alakították ki, hogy a Linux-rendszermagon csak a legcsekélyebb változtatásokat kelljen elvégezni.

A memóriakezelő felel a felhasználószintű alkalmazások kapcsolattartáshoz szükséges tárhelyeinek a kialakításáért. A memóriakezelő alrendszert úgy tervezték, hogy

1. minél kevesebb DMA-kérésre és -visszaigazolásra legyen szükség,
2. a gazdagépen minél kevesebb megszakításkérelemre legyen szükség,
3. elkerülje a költséges fizikai-virtuális címátalakító folyamatokat,
4. elkerülje a futásközbeni dinamikus memóriatúltelítődéseket.

Az OSM a DMA-adatok tárolására két különböző adatterületet tart fenn. A küldésre szánt adatok adatterületét az iNIC felé, és a fogadott adatok adatterületét – ezeket a foglalati eszköz küldi a rendszermag felé.

Ahogyan a *3. ábrán* látható, a Linux hálózati összetevői rétegelt szerkezetbe vannak szervezve. A felhasználói területen programozók a hálózati szolgáltatásokat foglalatokon keresztül érik el, felhasználva a Linux által biztosított foglalati rétegszolgáltatásokat. A ***linux/net.h*** fejlécfájlban kialakított foglalati szerkezet képezi a foglalati kapcsolati réteg alapját. A felhasználói szint alatt található az INET foglalati réteg, amely az IP-alapú protokollok (mint például TCP vagy UDP) közti kapcsolati csatlakozópontokat képezi. A réteg felépítését a ***net/sock.h*** fejlécfájlban található adatszerkezet-foglalat határozza meg.

Az INET foglalati réteg alatti réteget a foglalat típusa szabja meg, esetleg a TCP- vagy UDP-réteg, vagy közvetlenül az IP-réteg. Az IP-réteg alatt találhatók a hálózati eszközök, amelyek a csomagokat közvetlenül az IP-rétegtől kapják.

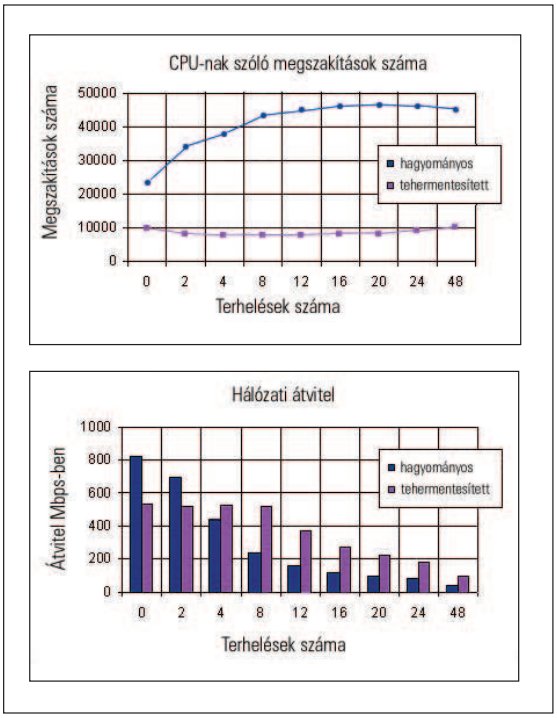
A foglalati OSM cseréli le az INET foglalati réteget. A foglalatok felől érkező összes foglalatokkal kapcsolatos üzenet I2O-üzenetekké alakítódik át, melyek végül az IOP-on található ISM felé továbbítódnak.

A beágyazott célpont

A beágyazott rendszerprogram a következő rétegekből épül fel: üzenetközvetítő réteg, TCP/IP-verem, eszközmeghajtó és RTOS. A program az üzenetközvetítő rétegnek azt a részét képezi, mely az OSM felől fogadja az üzeneteket, majd értelmezi őket, végül a kéréseket a TCP/IP-verem felé továbbítja. Ez a réteg felelős az üzenetek fogadásáért is, és ez küldi vissza a megfelelő válaszokat az OSM-nek. A teljesítménynöveléséért, valamint az örökölt osztott eszközmeghajtó rendszer miatt keletkező késések következményeit csökkentendő az üzenetközvetítő réteg az üzeneteket kötegekbe rendezi és csővezetéken keresztül továbbítja, egyúttal a válaszadásért is felel.

A TCP/IP-vermet teljes egészében a BSD 4.2 veremnek megfelelően építették fel, biztosítva a hálózati verem funkcionalitását

az üzenetközvetítő réteg felé. Mint az összes IOP-on futó egyéb program, a verem is az Intel 80310-es I/O processzor lapkakészletére van testreszabva, mely az Intel Xscale mikroszerkezetére épül (mely már fel van készítve az ARM-szerkezetre). A fejlesztés során a TCP/IP-vermet sokszor állították próbák elé, hogy a legkülönbözőbb mértékű hálózati forgalmat is a lehető leghatékonyabban kezelje.
A HDM-et úgy alakították ki, hogy a NIC-kel való áthárításból kifolyólag a lehető legtöbb haszonra tehessen szert. Ez magában foglalja a TCP- és IP-csomagok ellenőrzőösszegeinek ellenőrzését, az 1500 bájtnál nagyobb TCP-csomagok feldarabolását, és a lapka által támogatott megszakítások kötegelését. A következő NIC-lapkákhoz létezik támogatás: Intel 82550 Fast Ethernet Multifunction PCI/CardBus vezérlő, és Intel 82543GC Gigabit ethernetvezérlő.
Az RTOS olyan szabadalmaztatott operációs rendszer, melyet a bonyolult I/O-műveletek igényének megfelelően alakítottak ki. Ez az operációs rendszer teljes egészében I2O-képes. Ez részben azért alakult így, mert a tervezők egy prototípust is létre akartak hozni.
Mint már említettük, az alkalmazási réteg által kiadott foglalati hívások először átalakítódnak, majd pedig a PCI-sínen (bus) keresztül az I/O processzor felé továbbítódnak. Ez a beágyazott rendszer egy teljes értékű, I/O-átvitelek kezelésére kialakított számítógép, mely egy processzorból, memóriából, az RTOS-ból, valamint a PCI-sínből áll. Mivel elsősorban I/O-műveletekre hegyezték ki, nagyban lecsökkenti az üzenetváltások következményeit. Amint egy üzenet az IOP-hoz kerül, azonnal értelmezi. Az alkalmazás által kért foglalati hívás menten a beágyazott hálózati verem felé továbbítódik, és ha a foglalati művelet véget ér, az OSM rögvest értesítést kap.



40. ábra A próbák eredményei

A próbák eredménye

A prototípuson elvégzett teljesítménypróbák egyértelműen azt mutatják, hogy a TCP/IP-verem áthárításával mind a processzor terhelése, mind a gazdagép felé irányuló megszakításkérelmek száma csökkent. Egy hálózati szempontból teljesen leterhelt tesztgépen az áthárított verem képes volt a teljes forgalmat kezelni, a CPU pedig minden idejét az elsődleges alkalmazásoknak szentelhette. Egy áthárított veremmel nem rendelkező alaprendszeren a CPU a tevékenységét kénytelen volt rendszeresen megszakítani, de még így sem tudta a hálózati forgalmat teljes sebességgel ellátni, ami természetesen a hálózati forgalom lelassulásában mutatkozott meg.

Elképzelések a jövőt illetően

Az iSCSI (adattárolás IP-protokollon keresztül SCSI-csomagok TCP/IP-be csomagolásával) elterjedésével egyre nagyobb igény mutatkozik a hálózati forgalom növekedéséből adódó terhelés csökkentésére. A TCP/IP-verem IOP-ra való átültetésével az iSCSI-adapterek teljes értékű TCP/IP-verem támogatással fognak rendelkezni. Ha az iSCSI-t a normál SCSI programozói felület részévé tennék, az nagyban csökkentené a Linux-felületre kifejtett hatását. Ahhoz, hogy az iSCSI felvehesse a versenyt a Fibre Channel módszerrel, legalább hasonló teljesítményt kell tudnia felmutatni.
Ugyancsak a jövőre vonatkozó terv az is, hogy RTOS-ként Linuxot használjanak. A prototípus elkészítésekor egy Intel i960 RM/RN processzort használtak, de ekkor még nem állt rendelkezésükre beágyazott Linux. Azóta bemutatták az Intel Xscale mikroszerkezetet, amely a Linux alkalmazását a StrongARM-magon lehetővé tette. A Linux-alapú StrongARM Linux portolása az Xscale mikroarchitektúrára ez év végére befejeződik.
E mögött a prototípuskészítési terv mögött számos más cél is meghúzódott:

1. megmutatni, hogy a hálózati feladatoktól mentesített gazdagépprocesszor a hálózati forgalom elemzésére jóval kevesebb órajelciklust pazarol el;
2. bebizonyítani, hogy egy egyedi programmal felvértezett iNIC ugyanazon hálózati tevékenységek elvégzése mellett a hálózat teljesítményét maximálisan szinten képes tartani;
3. engedélyezni olyan I/O-processzorok használatát, amelyek a hálózati forgalom kezelésében képesek együttműködni a gazdagép processzorával, így maximalizálva csekély költséggel a Linux-kiszolgáló teljesítményét.

A TCP/IP-verem hálózati programokból álló környezetre történő áthárításának módszere beágyazott processzorok segítségével a teljesítménynövelés egyik leghatékonyabb módja. A nagysebességű hálózatok fejlődésében és a hálózati adattárolás elterjedésében a TCP/IP elkerülhetetlenül igen fontos szerepet fog játszani.

Technikai szakértők:
Dave Jiang, Dan Thompson, Jeff Curry, Sharon Bartmans, Don Harbin and Scott Goble.

Chen Chen
a fejlett I/O-alkalmazások terén végez kutatásokat az Intelnél. Elérhető a chen.chen@intel.com címen.

David Griego
több mint három éve lelkes Linux-rajongó. Szintén az Intelnél dolgozik, a tervezéstechnikai fejlesztési részlegnél. Elérhető a david.a.friego@intel.com címen