

Háskóli Íslands
Verkfræðideild

08.71.31 Gagnasafnsfræði
Jóhann Pétur Malmquist



HÁSKÓLI ÍSLANDS

Samhliða vinnsla fjölkjarna
örgjörva í gagnasöfnum

27. nóvember 2006
Sævar Öfjörð Magnússon

Efnisyfirlit

Inngangur	1
Kynning	2
TLS (Thread-Level Speculation)	3
TLS með færslum	4
Útrýming gagnahæðis	5
Mæliniðurstöður	6
Niðurstöður verkefnis	7
Um höfunda	8
Todd C. Mowry	8
J. Gregory Steffan	8

Inngangur

Við lifum svo sannarlega á spennandi tímum hvað framþróun í tölvubúnaði varðar. Hið klassíska lögmál Moores hefur átt vel við aukningu í fjölda transistora á örflögum, en á síðustu árum hefur rutt sér til rúms á almennum markaði tækni sem talið er að geti gjörbreytt þróun á örgjörvamarkaðnum. Fjölkjarna örgjörvar gætu verið það sem koma skal í tölvuvinnslu framtíðarinnar. Nú eru tvíkjarna örgjörvar orðnir nokkuð algengir í nýjum heimilistölvum og nýlega voru kynntir til sögunnar örgjörvar með 4 og 8 kjarna. Intel hefur sett sér það markmið að framleiða 80 kjarna örgjörva á næstu árum.

Það er því nokkuð ljóst að vélbúnaðarþróunin er hröð og spennandi. Eini gallinn við þessa þróun er að hugbúnaðarmarkaðurinn hefur ekki þróast með vélbúnaðinum á þessu sviði. Til þess að tveir eða fleiri kjarnar í örgjörva nýtist sem best í þungri vinnslu þarf að gera ráð fyrir þeim í forritun hugbúnaðar.

Greinin sem fjallað er um í þessum stutta texta fjallar um verkefni á vegum annars vegar Carnegie Mellon háskólans í Bandaríkjunum og Toronto háskóla í Kanada, og hins vegar Rannsóknarstofu Intel í Pittsburgh í Bandaríkjunum. Markmið verkefnisins var að nýta svokallað *intra-transaction parallelism*, eða samhliða vinnslu innan færslna, til þess að minnka töf í gagnasafnskerfum. Greinin birtist árið 2005 á VLDB ráðstefnunni (e. *Very Large DataBases*).

Sjálfur hef ég mikinn áhuga á samhliða vinnslu örgjörva og fannst mér því við hæfi að tengja gagnasafnsfræðina við þetta svið.

Kynning

Í byrjun er rétt að taka fram að íslenska orðið færsla er hér notað um enska orðið *transaction*.

Í greininni er byrjað á að tala um öra þróun vélbúnaðar á sviði fjölkjarna örgjörva. Allir helstu örgjörvaframleiðendur hafa framleitt fjölkjarna örgjörva þegar greinin er skrifuð árið 2005, eða Intel, IBM, AMD og Sun. Það er edlilegt að nýta nýja tækni og það sem höfundar hafa áhuga á er að nýta hana í að hraða vinnslu gagnasafnskerfa. Í gagnasafnskerfum sem nota OLTP (Online Transaction Processing) hafa þræðir aðallega verið notaðir í að auka gegnumstreymi gagna. Fræðilega séð mætti einnig nota þræði til þess að minnka töf eða biðtíma.

Flest gagnasafnskerfi í dag nýta sér svokallað *intra-query parallelism* eða samhliða vinnslu innan fyrirspurna, sem eykur ekki hraða nema í fyrirspurnum sem taka langan tíma. Greinarhöfundar telja hins vegar að *intra-transaction parallelism* sé ekki notað, og það af góðri ástæðu. Í fyrsta lagi þarf að breyta kerfinu þannig að það styðji marga þræði fyrir hverja færslu. Bæta þarf við læsingum sem færsluþræðir nota sameiginlega sem flækir kerfið og veldur töfum í keyrslu. Í öðru lagi þarf að skipta færslum upp í samhliða þræði.

Greinarhöfundar nota TPC-C til þess að mæla hraða gagnasafnsins. TPC er skammstöfun fyrir *Transaction Processing Council* og miðar að því að útbúa staðlaðar fyrirspurnir og gögn sem hægt er að nota til þess að mæla hraða og getu gagnasafnskerfa. Í TPC-C eru 5 meginfærslur. Greinarhöfundar nefna sem dæmi að í NEW ORDER færslunni (sjá mynd 1) er **for**-lykkja sem tekur 78% af keyrslutíma allrar færslunnar, og að slík lykkja sé kjörin til þess að sýna kosti þess að breyta færslum svo einstakir hlutir þeirra keyri samhliða. Þá er í raun hverri ítrun lykkjunnar breytt í stakan þráð, sem býður að sjálfsögðu upp á ákveðin vandamál. Það þarf að átt sig á því hvernig þræðirnir eru háðir innbyrðis, einnig verður að bæta við lásam til þess að við getum fullvissað okkur um að einungis einn þráður uppfæri gildi í töflu gagnagrunnsins í einu. Að lokum þarf að prófa færsluna með breytingum til þess að ákvarða hvort samhliða keyrslan gefi sömu niðurstöðu og sú sem er

keyrð í einni röð. Þetta ferli þarf að endurtaka oft og mörgum sinnum þar til fullnægjandi niðurstaða fæst.

```
begin_transaction {
  Read customer info [customer,warehouse]
  Read & increment order # [district]
  Create new order [orders, neworder]
  for ( each item in order ) {
    Get item info [item]
    if ( invalid item )
      abort_transaction
    Read item quantity from stock [stock]
    Decrement item quantity
    Record new item quantity [stock]
    Compute price
    Record order info [order_line]
  }
} end_transaction
```

Mynd 1: NEW ORDER færslan. Töflur sem eru notaðar í hverri línu eru innan hornklofa.

TLS (Thread-Level Speculation)

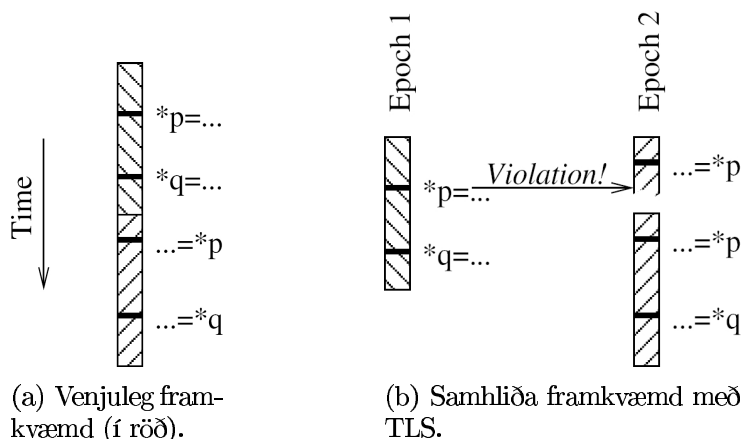
Það vandamál sem lýst hefur verið hér að framan þarf ekki endilega að vera mjög flókið í útfærslu. Tölvuhönnuðir hafa samhliða þróun fjölkjarna örgjörva útfært vélbúnaðarstuðning fyrir fjölkjarna gjörva sem kallaður er *Thread Level Speculation* (TLS). Þessi tækni er talin vera líkleg til þess að vera staðalbúnaður í fjölkjarna örgjörvum í framtíðinni, og hefur nú þegar verið útfærð í nokkrum gjörvum.

Í stuttu máli gerir TLS forritaranum kleyft að skipta forriti upp í nokkurs konar „hugsandi þræði“ (e. *speculative threads*) eða tímаскеið (e. *epochs*) sem eru framkvæmd samhliða. TLS tryggir að samhliða keyrsla gefi nákvæmlega sömu niðurstöður og hin upprunalega raðkeyrsla. Fylgst er með gagnahæði (e. *data dependency*) milli tímаскеiða og eru þræðir endurræstir ef framkvæmd þeirra vîkur frá hinni upprunalegu raðkeyrslu. Það að skipta forriti upp í slík tímаскеið eykur afköst án þess að breyta lokaniðurstöðunni.

Það má líkja framkvæmd einstakra tímаскеiða við framkvæmd svokallaðra „bjartsýnna“ færslna (e. *optimistic transactions*). Tímаскеiðin eru í eðli sínu mjög bjartsýn

og gera ráð fyrir að þau muni ekki skarast við önnur tímaskeið. Tímaskeiðin bera allar gagnafærslur saman við gagnafærslur fyrri tímaskeiða til þess að tryggja rétta útkomu og skrifast að lokum í minni þegar allt er í lagi en hætta við eða endurræsa þegar t.d. einhvers konar gagnahæði verður.

Greinarhöfundar mældu TPC-C færslur á BerkleyDB gagnasafnskerfinu eftir að hafa gert ákveðnar breytingar á því. Fyrst breyttu þeir byggingu gagnagrinda og aðferða sem allar færslur nota, svo hægt væri að útfæra samhliða keyrslu í hvaða færslu sem er. Einnig breyttu þeir grunneiningum sem öll gagnasöfn nota (svo sem lásur) sem þýðir að breytingarnar eru ekki endilega háðar BerkleyDB, og væri auðvelt að útfæra þær í öðrum gagnasöfnum í framhaldinu. Greinarhöfundar þurftu einungis að breyta um 1200 línur af 180,000 í forritskóða BerkleyDB. Það tók mastersnema um mánuð að útfæra breytingarnar. Mælingar leiddu svo í ljós að hægt var að eyða 36 til 76 % af biðtíma úr þremur af fimm TPC-C færslum.



Mynd 2: TLS passar að allur lestur og skrif gerist í réttri röð.

TLS með færslum

TLS leyfir forritaranum að búa til samhliða þræði eða tímaskeið, án þess að þurfa að hugsa um hvort gögnin komi rétt út á endanum eða ekki. TLS sér alfarið um að tryggja það. Sem dæmi nefna greinarhöfundar færslu sem uppfærir línur í töflu. Ef línurnar eru fullkomlega óháðar, er auðvelt að sjá að þær má uppfæra samhliða. Í raun og veru eru raðnúmer línanna ekki vituð fyrirfram svo þær þarf að uppfæra hverja á eftir annarri til

Þess að varðveita ástönd sem myndast milli uppfærslna. Með TLS má setja þetta upp sem samhliða vinnslu, sem takmarkast eingöngu af milliniðurstöðum sem mynda gagnahæði.

Útrýming gagnahæðis

Greinarhöfundar útfærðu TLS í NEW ORDER færslunni úr TPC-C benchmark staðlinum á BerkleyDB. NEW ORDER var valin sem sýnidæmi þar sem mikið er af les- og skrifaðgerðum, hún stendur fyrir helming vinnuálagsins, og hún er notuð til þess að mæla gagnaflæði í TPC-C. Sauðakóða færslunnar má sjá á mynd 1. Takmark höfunda var að útfæra samhliða vinnslu í **for**-lykkjunni sem stendur fyrir 78% vinnuálags færslunnar. Lykkjan les `item` töfluna, uppfærir `stock` töfluna og bætir við `order_line` töfluna. Lestur úr `item` töflunni getur ekki valdið gagnahæði milli tveggja tímaskeiða. Uppfærsla `order_line` töflunnar ætti ekki heldur að valda gagnahæði þar sem ný lína er búin til fyrir hvert tímaskeið. Uppfærsla á `stock` töflunni er hugsanlegt vandamál – Ef tvö tímaskeið vísa í sama hlutinn, gæti uppfærsla í einu tímaskeiði valdið vandræðum í öðru. Í rauninni eru líkindi á slíku gagnahæði mjög lítil og er þessi lykkja því kjörinn kandídat til samhliða vinnslu.

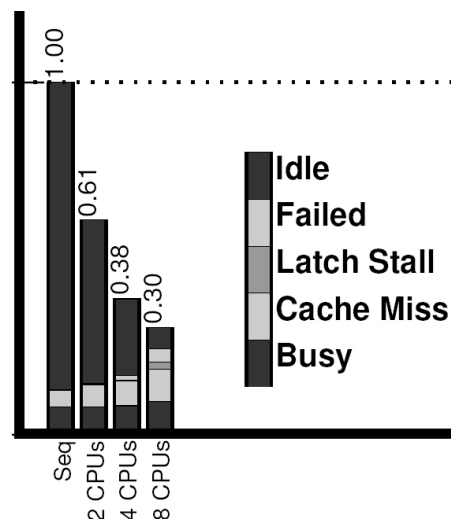
Þegar kom að því að útfæra breytingarnar kom í ljós gagnahæði á mun lægra plani heldur en í lýsingunni á færslunni hér að framan: Lestur/skrif í læsingar, biðminni, atburðaskrá og B-tré. Greinarhöfundar leggja til þrjár aðferðir til þess að losna við slíkt gagnahæði:

1. **Uppskipting gagnagrinda.** Aðferð sem tekur frá minni fyrir gagnagrunninn, `db_malloc`, tekur venjulega frá einn haug af minni sem er notaður í allt. Samhliða aðgangur í slíkt minni er ómögulegur og veldur árekstrum. Taka þarf frá sérstakt minni fyrir hvert tímaskeið til þess að losna við þá. Ef upp koma fleiri tilfelli þar sem tímaskeið deila minni þarf að skipta því upp sömuleiðis.
2. **Nýting einangraðra afturkallanlegra aðgerða (IUO).** TLS tryggir að þegar minnisblokk er sótt og fest (`pin_page`) í biðminni af einu tímaskeiði þá klárast sú aðgerð áður en annað tímaskeið getur kalla á `pin_page`. Greinarhöfundar vilja frek-

ar sleppa því að nota TLS þegar kemur að þessari aðgerð, þar sem `pin_page` er svokölluð einangruð og afturkallanleg aðgerð. Einangraða aðgerð má afturkalla án þess að það hafi áhrif á aðra þræði eða tímaskeið. Því eru einangraðar afturkallanlegar aðgerðir keyrðar án þess að hugsa um gagnahæði því auðvelt er að afturkalla þær ef tímaskeiðið hættir við eða endurræsir.

3. **Fresta aðgerðum þar í lok tímaskeiða.** Þegar lína í atburðaskrá er færð inn, fær hún raðnúmer og ákveðinn teljari er hækkaður. Þessi teljari myndar gagnahæði milli tveggja tímaskeiða. Greinarhöfundar leggja til innsæi sitt og sjá að tímaskeiðin nota í raun aldrei raðnúmerin sjálf, heldur búa þau eingöngu til. Færslur í atburðaskrána má búa til á meðan tímaskeiðin eru framkvæmd, en þær fá ekki raðnúmer fyrr en í lokin, þegar síðasta tímaskeiðið hefur hætt keyrslu. Almennt séð, þegar aðgerð hefur ekki áhrif á keyrslu tímaskeiðsins sjálfs, heldur á aðrar færslur, þá má seinka henni þar til í lok tímaskeiðs.

Mæliniðurstöður



Mynd 3: Mæliniðurstöður fyrir NEW ORDER færsluna.

Mynd 3 sýnir ágætlega hvernig afköst jukust og biðtími minnkaði. Á lóðrétta ásnum er keyrslutími. Súlan sem er lengst til vinstri er hin upprunalega raðkeyrsla og svo koma súlur fyrir 2, 4 og 8 kjarna örgjörva.

Niðurstöður verkefnis

Greinahöfundar meta í lokin hversu vel heppnast til með breytingarnar. Gagnagrunnasamfélagið hefur ávallt notað samhliða vinnslu í gegnum tíðina og það er ekki vafi á því að það er bara tímaspursmál þar til gagnagrunnskerfi nýta sér þræði og fjölkjarna örgjörva til þess að auka vinnslugetu sína. Tölvuhönnuðir hafa eins og áður sagði komið með tól sem auðveldar til muna breytingar á hugbúnaði sem þegar er til staðar – *Thread Level Speculation* (TLS). TLS var upprunalega hannað til þess að auðvela forriturum að aðlaga venjuleg notendaforrit að samhliða vinnslu, en gerir gagnagrunnhönnuðum nú kleift að nýta samhliða vinnslu sem áður var óþekkt.

Greinahöfundar beina athygli sinni að einu slíku tækifæri sem TLS býður upp á, en það er að koma á samhliða vinnslu innan færslanna sjálfra. Verkefnið sýnir fram á að hægt er að ekki einungis auka gagnastreymi, heldur minnka biðtíma þriggja af fimm færslum TPC-C um 36-74% með því að nota TLS á gjörva með fjóra kjarna, eða um 29-44% með tveimur kjörnum. Fyrri aðferðir hafa lagt áherslu á að forritarinn sjálfur útfæri samhliða vinnsluna, en hér er komin aðferð sem minnkar þá vinnu niður í að skilgreina tímaskeið sem keyra skuli samhliða. Í framtíðinni myndi færsluþýðandi (e. *transaction compiler*) sjá um þá vinnu að mestu. Greinahöfundar sýndu að hægt var að breyta gagnasafnskerfi á þennan hátt, og þá vinnu er hægt að heimfæra á flest gagnasafnskerfi sem eru á markaði í dag.

Um höfunda

Todd C. Mowry

Todd C. Mowry er fæddur árið 1966. Hann útskrifaðist með mastersgráðu í rafmagnsverkfræði og doktorsgráðu frá Stanford háskóla 1989 og 1994. Hann var aðstoðarprófessor í tölvuvísindum við Toronto háskóla 1994-1997 áður en hann flutti sig til Carnegie Mellon háskóla þar sem hann hefur verið prófessor. Hann hefur tekið sér hlé frá kennslu og í dag er hann forstöðumaður rannsóknastofu Intel í Pittsburgh og stundar rannsóknir og þróun á fjölkjarna örgjörvum og tækni sem tengist þeim. Mowry hefur verið talinn með fremstu tölvuhönnuðum heims og kynnti t.d. fyrstur tækni í örgjörvum sem sækir gögn án þeirra er þörf (e. *prefetch*). Hans helstu markmið eru að þróa nýjar aðferðir í vélbúnaði sem og hugbúnaði til þess að auka afköst án þess að auka álagið á forritaranum.



J. Gregory Steffan

Greg Steffan var doktorsnemi í tölvuvísindadeild Carnegie Mellon háskóla. Þar var hann meðlimur í STAMPede verkefni sem áður nefndur Todd Mowry stýrði. Hann hlaut BS og MS gráður í tölvuverkfræði frá Toronto háskóla 1995 og 1997. Hann hefur unnið fyrir MIPS Technologies Inc. og ALPHA þróunarhópi Compaq. Hann er nú aðstoðarprófessor í tölvuverkfræði við Toronto háskóla þar sem hann stundar rannsóknir á sviði tölvuhönnunar, hugbúnaðarþýðenda, samhliða og dreifðum kerfum, líkanagerð og hermun. Hann spilar á klassískan gítar, stundar golf og skvass og nýtur þess eindregið að lýsa sjálfum sér í þriðju persónu, eins og hann segir sjálfur frá.



Samhliða vinnsla fjölkjarna örgjörva í gagnasöfnum

Sævar Öfjörð Magnússon (som1@hi.is)
30. nóvember 2006

1

Samhliða vinnsla í gagnasöfnum

- Mikil þróun í fjölkjarna örgjörvum, þ.e. tæknilega hliðin.
- Hugbúnaður þarf að fylgja tækninni eftir og gera ráð fyrir möguleikum á samhliða vinnslu, þ.m.t. gagnasöfn.
- OLTP gagnasafnskerfi hafa notað þræði í að auka flæði gagna.
- Greinarhöfundar vilja nota þræði til þess að minnka biðtíma eða tóf og þar af leiðandi auka afköst.
- Gagnasafnskerfi hafa notað *intra-query parallelism*, greinarhöfundar vilja innleiða *intra-transaction parallelism*.

2

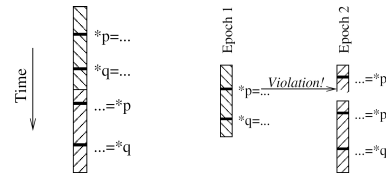
Samhliða vinnsla í gagnasöfnum

- Nota TPC-C staðlaðar færslur/fyrirspurnir og gagnasöfn til þess að mæla afköst.
- Ein færslan er kölluð NEW ORDER – inniheldur for-lykkju sem tekur 78% keyrslutímans.
- Kjörin færsla til samhliða vinnslu.
- Í NEW ORDER er hverri ifrun lykkjunnar breytt í þráð eða tímaskeið (e. *epoch*).
- Byður upp á vandamál, t.d.
 - Þræðir gætu verið háðir innbyrðis.
 - Læsa þarf gögnum milli þráða.

3

TLS (Thread Level Speculation)

- Tækni sem hefur þróast samhliða fjölkjarna örgjörvum.
- Gerir forritun fyrir samhliða vinnslu einfaldari.
- Forritarinn þarf að skipta forriti upp í þræði (*speculative threads*) eða tímaskeið sem eru keyrð samhliða.
- Ef gagnahæði kemur upp eru þræðir einfaldlega endurræstir.



4

TLS

- TLS tryggir að samhliða keyrsla gefi nákvæmlega sömu niðurstöður og hin upprunalega raðkeyrsla.
- Tímaskeiðin eru látin vera „bjartsýn“ (e. *optimistic*).
- Gera ráð fyrir að þau muni ekki skarast við önnur tímaskeið.
- Gagnafærslur milli tímaskeiða eru bormar saman og ef misræmi verður er þráður endurræstur.
- Forritarinn getur því óhræddur skipt kóðanum upp í tímaskeið – þarf ekki að hafa áhyggjur af lokaniðurstöðu.
- Skipting í tímaskeið ytri hlutverk þýðanda.

5

Útfærsla og útrýming gagnahæðis

- NEW ORDER valin vegna mikils fjölda les- og skrifaðgerða – notuð til að mæla gagnafæði í TPC-C.
- Ýmis konar gagnahæði komu upp, og þá aðallega frekar djúpt í byggingu kerfisins.
- Lestur/skrif í læsingar, biðminni, atburðaskrár og B-tré.
- Greinarhöfundar lýsa nokkrum aðferðum til þess að útrýma gagnahæði.

6

Útrýming gagnahæðis – Uppskipting gagnagrinda

- `db_malloc` tekur frá minni fyrir gagnagrunninn, og þá yfirleitt eitt svæði fyrir allt ⇒ samhliða aðgangur ómögulegur.
- Taka þarf frá sérstakt minni fyrir hvert tímaskeið eða þráð.
- Ef upp koma tilfelli þar sem tímaskeið deila minni þarf að skipta því upp sömuleiðis.

7

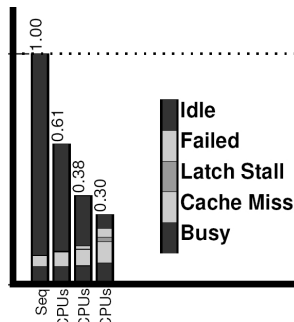
Frestun aðgerða þar til í lok tímaskeiða

- Þegar lína í atburðaskrá er færð inn fær hún raðnúmer og teljari er hækkaður.
- Teljarinn myndar gagnahæði milli tímaskeiða.
- Tímaskeiðin nota aldrei raðnúmerin sjálf, heldur búa þau til.
- Færslur í atburðaskrá má búa til við framkvæmd, en fá ekki raðnúmer fyrr en síðasta tímaskeið hefur lokið keyrslu.
- Almennt má seinka aðgerð þar til í lokin ef hún hefur ekki áhrif á keyrslu tímaskeiðsins sjálfs.

8

Mæliniðurstöður

- Mynd sýnir hvernig keyrslutími breyttist eftir fjölda kjarna.



9

Niðurstöður

- Samhliða vinnsla hefur ávallt verið notuð í gagnasafnskerfum.
- Greinin bendir á möguleika sem hefur ekki verið hendi fyrir en með auknum vélbúnaðarstuðningi.
- TLS er tól sem auðveldar breytingar á gagnasafnskerfum.
- Notaður var BerkleyDB. Innan við 1200 línur af 180,000 þurfti að breyta í kóðamum. Verkið tók mánuð.
- Aðallega breytingar á aðferðum sem eiga við um allar færslur svo og breytingar á grunnagnagrindum gagnasafnskerfa, svo sem láusum, sem ætti að vera auðvelt að yfirfæra á flest kerfi.

10

Höfundar: Todd C. Mowry

- Fæddur 1966. MS í rafnagnsverkfræði 1989 og doktorsgráða 1994 frá Stanford.
- Aðstoðarprófessor við Toronto háskóla 1994-1997, flutti sig til Carnegie Mellon þar sem hann hefur verið prófessor.
- Í dag – forstöðumaður rannsóknastofu Intel í Pittsburgh, stundar rannsóknir og þróun á fjölkjarna örgjörvum og tækni sem tengist þeim.
- Upphafsmáður *prefetch* tækninnar.



11

Höfundar: Gregory Steffan

- BS og MS frá Torontoháskóla í Kanada 1995 og 1997. Doktorsgráða frá Carnegie Mellon og vann þar með Todd Mowry.
- Hefur unnið fyrir MIPS og verið í ALPHA þróunarhópi Compaq.
- Er nú aðstoðarprófessor við Toronto háskóla þar sem hann stundar rannsóknir á sviði tölvuhönnunar, hugbúnaðarþýðenda o.fl.



12

Spurningar?

13

Optimistic Intra-Transaction Parallelism on Chip Multiprocessors

Christopher B. Colohan*, Anastassia Ailamaki*, J. Gregory Steffan†, and Todd C. Mowry*‡

*School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
USA
{colohan,natassa,tcm}@cs.cmu.edu

†Department of Electrical &
Computer Engineering
University of Toronto
Toronto, Ontario M5S 3G4
Canada
steffan@eecg.toronto.edu

‡Intel Research Pittsburgh
4720 Forbes Ave., Suite 410
Pittsburgh, PA 15213
USA
todd.mowry@intel.com

Abstract

With the advent of chip multiprocessors, exploiting intra-transaction parallelism is an attractive way of improving transaction performance. However, exploiting intra-transaction parallelism in existing database systems is difficult, for two reasons: first, significant changes are required to avoid races or conflicts within the DBMS, and second, adding threads to transactions requires a high level of sophistication from transaction programmers. In this paper we show how dividing a transaction into *speculative threads* solves both problems—it minimizes the changes required to the DBMS, and the details of parallelization are hidden from the transaction programmer. Our technique requires a limited number of small, localized changes to a subset of the low-level data structures in the DBMS. Through this method of parallelizing transactions we can dramatically improve performance: on a simulated 4-processor chip-multiprocessor, we improve the response time by 36–74% for three of the five TPC-C transactions.

1 Introduction

We are in the midst of a revolution in microprocessor design: computer systems from all of the major man-

ufacturers that feature chip multiprocessors (CMPs) and simultaneous multithreading (SMT) are entering the marketplace. Examples include Intel’s “Smithfield” (dual-core Pentium IV’s with 2-way SMT), IBM’s Power 5 (combinable, dual-core, 2-way SMT processors), AMD’s Opteron (dual-core), and Sun Microsystems’s Niagara (an 8-processor CMP). How can database systems exploit this increasing abundance of hardware-supported threads? Currently, for OLTP workloads, threads are primarily used to increase transaction *throughput*; ideally, we could also use these parallel resources to decrease transaction *latency*. Although most commercial database systems do exploit *intra-query parallelism* within a transaction, this form of parallelism is only useful for long running queries, while OLTP workloads tend to issue multiple short queries. To the best of our knowledge, commercial database systems do not exploit *intra-transaction parallelism* [2, 11, 21], and for good reason.

Parallelizing a transaction is difficult. First, the DBMS must be modified to support multiple threads per transaction. Latches must be added to data structures which are shared between threads in the transaction. These latches add complexity and hinder performance. Second, the transaction must be divided into parallel threads. Consider the NEW ORDER transaction, which is the prevalent transaction in TPC-C [6] (Figure 1). We can parallelize the main loop (which represents 78% of the execution time), such that each loop iteration runs as a thread. The transaction programmer must understand when these threads may interfere with each other, and add inter-thread locks to avoid problems; e.g., the thread should use inter-thread locks to ensure that only one thread updates the quantity of an item in the `stock` table at a time. Finally, the transaction programmer must test the new transaction to ensure that the resulting parallel execution is correct and ensure that no new deadlock conditions or subtle race conditions were introduced, and

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the VLDB copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Very Large Data Base Endowment. To copy otherwise, or to republish, requires a fee and/or special permission from the Endowment.

**Proceedings of the 31st VLDB Conference,
Trondheim, Norway, 2005**

```

begin_transaction {
  Read customer info [customer, warehouse]
  Read & increment order # [district]
  Create new order [orders, neworder]
  for(each item in order){
    Get item info [item]
    if(invalid item)
      abort_transaction
    Read item quantity from stock [stock]
    Decrement item quantity
    Record new item quantity [stock]
    Compute price
    Record order info [order_line]
  }
}
end_transaction

```

78% of transaction execution time

Figure 1: The NEW ORDER transaction. In brackets are the database tables touched by each operation.

then repeat the entire process until satisfactory performance is achieved.

1.1 Incremental Parallelization with Thread-Level Speculation

Fortunately, computer architecture researchers have anticipated this problem and have developed hardware support for chip multiprocessors which eases parallelization, called *Thread Level Speculation* (TLS) [7, 18, 19]. Support for TLS is simple and elegant, has been implemented in Sun’s MAJC [19] and Stanford’s Hydra [7] CMPs, and is a strong candidate technology for future high-end CMPs. In a nutshell, TLS allows a program to be divided arbitrarily into speculative threads (or *epochs*) which are executed in parallel. The TLS mechanism ensures that the parallel execution is *identical* to the original sequential execution. It preserves sequential semantics by tracking data dependencies between epochs and restarting threads when their execution diverges from the original sequential execution (TLS is described in more detail in Section 2). In essence, dividing a program into epochs improves performance without affecting correctness.

Our goal is to parallelize important OLTP transactions, to achieve high performance with low implementation overhead. In this study we parallelize the main loops of the transactions from TPC-C, running on the BerkeleyDB DBMS [13]. We execute that transaction on a TLS system which provides profile information identifying the performance bottleneck; we then perform localized optimizations on the DBMS code to remove that bottleneck. Removal of each bottleneck exposes the next bottleneck, so we repeat the process. This paper describes the bottlenecks we encountered (Section 4), and provides general techniques for eliminating them (Section 3).

While in this paper we evaluate TPC-C transactions running on BerkeleyDB, our techniques can be generalized in two important ways. First, the changes we made were to DBMS data structures and functions which are shared by all transactions, hence the optimizations we describe can be applied to any transaction. Second, we change fundamental primitives used by all database systems (such as latches and locks), hence our techniques are not specific to BerkeleyDB and can be applied to other database systems. Applying our techniques required changing less than 1200 lines out of 180,000 lines of code in the DBMS, and took a graduate student about one month of work. As a result we eliminate 36 to 74% of the latency from three out of five TPC-C transactions.

1.2 Related Work

Traditionally, high-performance database systems have targeted inter-transaction parallelism, or intra-operation parallelism, while this paper introduces new techniques for exploiting intra-transaction parallelism. Previous work on intra-transaction parallelism has focused on hand-parallelized transactions [9, 15]—requiring great effort from the programmer and significant changes to the DBMS. Shasha *et. al.* showed that if a conflict graph can be constructed for a set of transactions, then the transactions can be *chopped* into smaller transactions which increases the degree of concurrency in the workload [16]. In this paper we show how to exploit intra-transaction parallelism with very little effort by the transaction programmer, minimal changes to the DBMS, and without having to construct a complete conflict graph.

Our paper draws upon Kung and Robinson’s optimistic concurrency control work [10] in two ways. First, the execution of *epochs* in thread level speculation is very similar to the execution of optimistic transactions: epochs optimistically assume that they will not conflict with other epochs, epochs compare all of their reads and writes to earlier epochs to ensure a serializable execution, and epochs commit when they succeed or abort and restart when speculation fails. Second, we optimistically omit lock and latch acquires in epochs, and let the TLS mechanism resolve conflicts. Our technique for optimistically omitting locks is also similar to transactional memory [8]. The difference between our paper and both transactional memory and optimistic concurrency control is that we allow speculative transactions to interact with non-speculative transactions, with *no* changes to the non-speculative transactions.

The idea of using TLS to simplify manual parallelization was first proposed by Prabhu and Olukotun for parallelizing SPEC benchmarks on the Hydra chip multiprocessor [7, 14]. The base Hydra multiprocessor uses a design similar to the shared cache CPU design used in the evaluation portion of this paper.

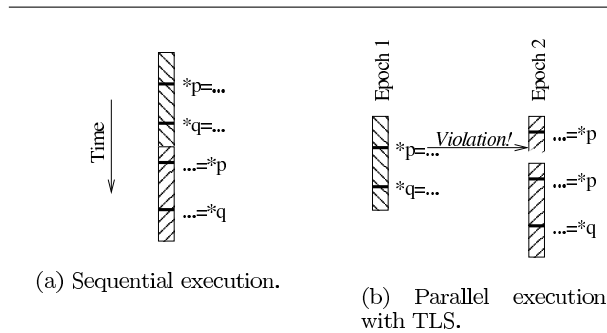


Figure 2: How TLS ensures that all reads and writes occur in the original sequential order.

1.3 Contributions

Our paper makes the following contributions: (i) it solves the problem of parallelizing the central loop of a transaction, which reduces transaction latency and hence decreases contention for resources used by the transaction; (ii) it provides a methodology for eliminating the data dependences which limit parallel performance, describing three specific techniques for eliminating these dependences and examples of their application; (iii) it demonstrates the application of these techniques by incrementally parallelizing a transaction running on a real DBMS, reducing transaction latency by more than a factor of two on a four-CPU machine.

2 Applying TLS to Transactions

When extracting intra-transaction parallelism, thread level speculation (TLS) allows the programmer to parallelize aggressively without worrying about correctness. Consider a transaction which updates several rows in a database: if the rows are indeed independent, then these updates could be performed in parallel. However, since the row IDs are typically not known in advance, the updates must instead be performed sequentially to preserve potential read and write dependences between different updates to the same row. With TLS support, the updates could be performed aggressively in parallel, limited only by the actual runtime dependences between rows. The following describes the basic functionality of TLS, including both software and hardware support.

2.1 Software Support for TLS

Under TLS, sequential code (Figure 2(a)) is divided into *epochs*, which are executed in parallel by the system (Figure 2(b)). The system is aware of the original sequential order of the epochs, and also observes every read and write to memory that the epoch performs (i.e. the reads and writes through *p* and *q*).

The system observes whether epoch 1 ever writes to a memory location which has already been read by epoch 2—if so, then epoch 2 has *violated* sequential

semantics, and is rewound and re-executed with the correct value. For example, in Figure 2(b) we see that epoch 2 read *p* before epoch 1 wrote to *p*, so we restart epoch 2. On the second execution epoch 2 reads the new value. Note that the read of *q* does not cause a violation, since it executes *after* the write to *q*, and thus reads the correct value. By observing all loads and stores, and restarting an epoch whenever it consumes an incorrect value, the TLS mechanism ensures that the parallel execution is identical to the original sequential execution.

The execution of epochs is similar to the execution of transactions in a system with optimistic concurrency control [10]: an epoch either commits or is violated (aborts), and if there are no dependences between epochs then their concurrent execution will be successful. The difference is that epochs are much smaller than transactions (we demonstrate the use of between 5 and 200 epochs per transaction in this paper by creating one epoch per loop iteration), and hardware support makes the cost of violating and restarting an epoch much lower than the cost of aborting and restarting a transaction. In addition, the ordering of the epochs is specified by their epoch number when the epochs are created, while with optimistic concurrency control the serial ordering is determined by the transaction commit order.

How does a programmer use this programming model to incrementally improve performance? First, and most importantly, the programmer can simply specify the decomposition of the transaction into epochs and do nothing further, and the result will be a correct parallel execution of the transaction (albeit one that does not necessarily perform very well). The epochs will likely have dependences between them which will cause violations (i.e., cause speculation to fail). If this is the case, then the programmer can use feedback from the TLS system to identify the cause of each dependence and try to eliminate them.

For example, one dependence which we encountered when parallelizing `NEW ORDER` was in the buffer pool: every time a page was requested and it was present in the buffer pool, the global variable `st_cache_hit` would be incremented. This increment would both read and write the variable, hence whenever an epoch requested a page from the buffer pool it would cause any later epochs which had already made a buffer pool request to be violated and restart. Once the system identified this dependence it was easy to correct—we changed the code so that there was one copy of `st_cache_hit` per CPU, and updated the appropriate copy of the variable on a buffer pool hit. We then modified the code which reads this variable (which is only invoked rarely, since this variable is used only as a performance monitoring statistic) to sum all of the per-CPU copies. Therefore a simple change to the code can eliminate a performance-limiting data depen-

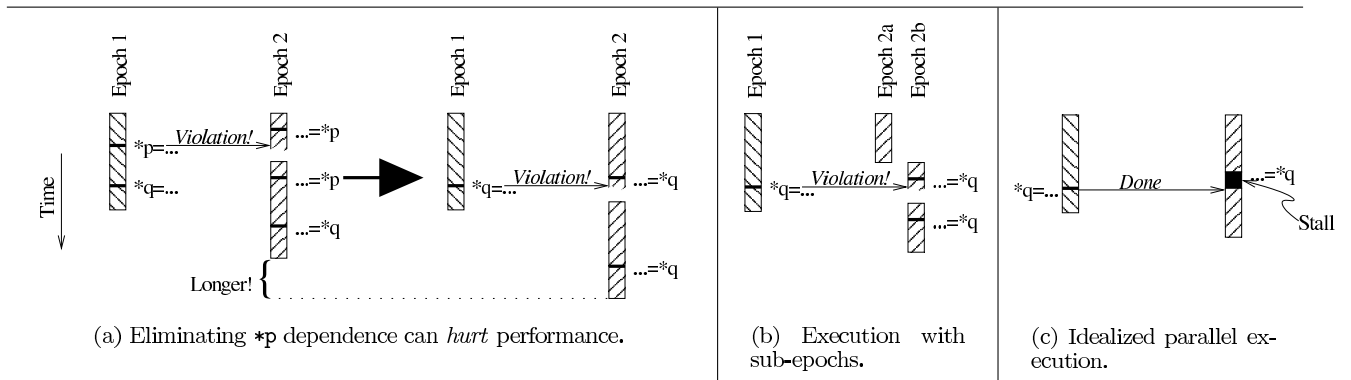


Figure 3: Sub-epochs improve performance when dependences exist.

dence.

2.1.1 Tolerating Dependence Violations

Previous TLS work has focused on programs with small epochs that have infrequent data dependences between them. Eliminating the only frequent data dependence between a pair of epochs removes the main source of violations, and results in a large performance gain. Unfortunately, in database systems the epochs are large and have many frequent data dependences between epochs. How does eliminating a data dependence affect these epochs?

In Figure 3(a) illustrates the result of eliminating the dependence caused by p . In this particular case, removing that dependence actually *hurts* performance! The problem is that the elimination of one dependence can expose other dependences, and merely delay an inevitable roll-back and re-execution. This problem can be avoided by dividing an epoch into several *sub-epochs* (Figure 3(b)). A sub-epoch acts like a checkpoint or sub-transaction: when a violation occurs, execution is rolled back just to the last sub-epoch so that less work has to be redone. Compare the execution with sub-epochs (Figure 3(b)) to an idealized parallel execution where each read does not execute until any dependent write has already executed (Figure 3(c)): as an epoch is broken into more and more sub-epochs, it approaches the performance of this idealized execution. Sub-epochs do not add parallelism, and hence have a lower hardware overhead than epochs. The hardware overhead of sub-epochs limits the number of supported sub-epochs [1], in this paper we assume 8 sub-epochs per epoch are used.

2.2 Hardware Support for TLS

The study presented in this paper is based on a hardware implementation of TLS which buffers speculative state in the cache hierarchy, and detects dependence violations by observing cache coherence messages [5, 18]. Previous TLS designs focused on applications with small epochs (50 to 5,000 dynamic in-

structions), while the large epochs (between 7,500 and 490,000 instructions) in a database transaction require the additional ability to buffer the state of very large epochs, as well as support for sub-epochs so that violations have a tolerable performance penalty. Concurrently with the work presented in this paper, we also designed a new implementation of TLS hardware that supports both large epochs and sub-epochs, the details of which are beyond the scope of this paper and are available in a technical report [1].

3 Eliminating Dependences

To evaluate the potential of applying TLS to a database system, we implemented the NEW ORDER transaction from the TPC-C benchmark on the BerkeleyDB storage manager. We chose NEW ORDER as the representative transaction in TPC-C because it is rich in read and update operations, it accounts for half of the workload, and it is used to measure throughput in TPC-C (TMPC). Pseudo-code for the transaction is shown in Figure 1. We want to parallelize the `for` loop in the transaction, as this loop comprises 78% of execution time. At a high level, this loop reads the `item` table, updates the `stock` table, and appends to the `order_line` table. Since it is read-only, the read of the `item` table cannot cause a data dependence between epochs. The append to the `order_line` table should not cause data dependences either, since a new order line is generated for each epoch. The update of the `stock` table is potentially problematic—if two epochs were to refer to the same item, then one epoch’s update of the `stock` table might interfere with another epoch’s update. However, in the benchmark specification items are randomly chosen uniformly from a set of 100,000 items, and so the probability of any two epochs conflicting is very small. At this high level, thanks to the infrequency of data dependences, it appears that this loop is an ideal candidate for TLS parallelization.

While our high-level analysis concludes that TLS parallelization is promising, the implementation details of query execution algorithms and access

methods reveal more potentially-limiting data dependences: read/write accesses to locks, latches, the buffer pool, logging, and B-tree indexes will cause data dependences between epochs. To eliminate these data dependences we propose and analyze three techniques:

1. Partition data structures. A memory allocation operation (`db_malloc`) typically uses a single pool of memory, hence parallel accesses to this shared pool will conflict. Using a separate pool of memory for each concurrent epoch avoids such conflicts. Many other dependences are also due to multiple epochs sharing a resource in memory—these dependences can be avoided by partitioning that resource.

2. Exploit isolated undoable operations (IUOs).

The TLS mechanism ensures that all attempts to fetch and pin a page (`pin_page`) in the buffer pool by one epoch complete before any invocations of `pin_page` in the next epoch begin, due to conflicts in the data structures which maintain LRU information. We prefer to allow `pin_page` operations to complete in any order. An epoch can simply call `pin_page` with speculation disabled: if the epoch is violated then the fetched page just remains in the buffer pool, and `unpin_page` can be invoked to release the page. This works because the `pin_page` operation is *undoable* and *isolated*. An undoable function has a corresponding function which can undo the isolated function's effects. An isolated function can be undone without affecting any other epoch or thread in the system. When speculatively parallelizing a transaction, we exploit isolated undoable operations by executing them non-speculatively, and call the corresponding undo function if the epoch is violated. This is similar to nested top actions in ARIES [12], since we modify the execution but preserve higher level semantics.

3. Postpone operations until the end of the epoch. When a log entry is generated, it is assigned a log sequence number and increments a global variable. This log sequence number counter forms a dependence between these two epochs. Our key insight was that an epoch never uses log sequence numbers—it only generates them. We can generate log entries during the execution of the epoch, and assign all of the sequence numbers at the end of the epoch after all previous epochs have completed, and just before committing the epoch (which makes the new log entries visible to the rest of the system). When an operation has no impact on the execution of the epoch, and instead only affects other transactions then it can be delayed until the end of the epoch.

In the next section we explore the major subsystems of the DBMS, and show how these three techniques can be used to eliminate the critical dependences we encountered while tuning the NEW ORDER transaction.

4 Performance Tuning

When we first parallelized the NEW ORDER transaction we encountered many dependences throughout the DBMS code. Some dependences are easy to eliminate through a local change to the source code: for example, false sharing dependences (see Section 4.5) can be eliminated by inserting padding in data structures so that independent variables do not share a single cache line. Other data dependences are inherent in the basic design of the database system, such as the creation of log sequence numbers or the locking subsystem. In the following sections we tour the database system's major components, and explain how we eliminate or avoid dependences on the common path in order to increase concurrency for TLS parallelization.

4.1 Resource Management

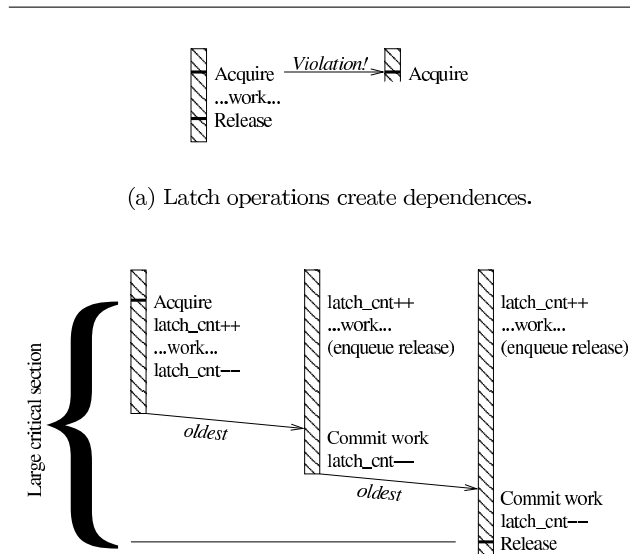
A large portion of every DBMS is concerned with the management of resources, including latches, locks, cursors, private and shared memory, and pages in the buffer pool. All of these resources can be acquired and released. Dependences between epochs occur when two epochs try to acquire the same resource, or when the data structures which track unused resources are shared between epochs. In the next sections we examine each of these resources and develop strategies for executing them in parallel.

4.1.1 Latches

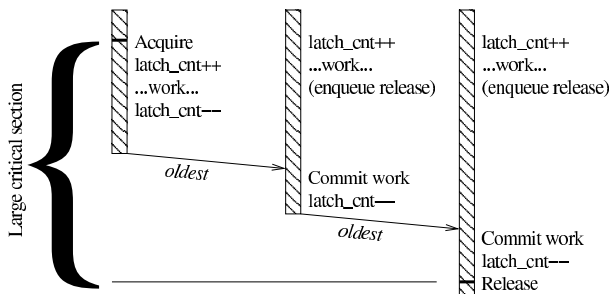
The database system uses latches extensively to protect data structures, and as a building block for locks. Latches are required for correct execution when multiple transactions are executing concurrently, and ensure that only one thread is accessing a given shared data structure at any time. Latches are typically held only briefly—in Section 4.1.2 we discuss *locks*, which offer concurrency control for database entities.

Latches form a dependence between epochs because of how they are implemented: a typical implementation uses a read-test-write cycle on a memory location (which may be implemented as a test-and-set, load-linked/store-conditional, atomic increment, etc.). This read-test-write cycle can cause a data dependence violation between epochs (Figure 4(a)).

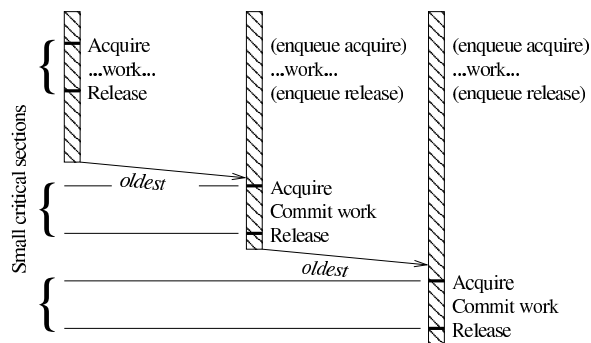
Under TLS, using latches to serialize accesses *within a transaction* is completely unnecessary—the TLS mechanism already ensures that any data protected by the latch is accessed in a serializable order, namely the original sequential program order. Hence using latches to preserve mutual exclusion between epochs is redundant with the existing TLS mechanism [8]. However, latches do ensure that mutual exclusion is maintained *between transactions*, and TLS does not perform that function. So we cannot simply discard the latches; we must instead ensure that they preserve mutual exclusion between transactions without causing violations



(a) Latch operations create dependencies.



(b) Aggressive latch acquire. The long critical section that results may cause performance issues.



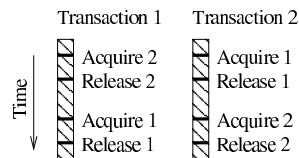
(c) Lazy latch acquire. Delaying the acquire shrinks the critical section.

Figure 4: Adapting latches for use under TLS execution.

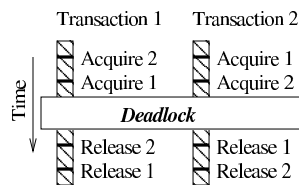
between the epochs within a transaction.

There are two operations performed on a latch: *acquire* and *release*. Let us first consider release operations. When a latch is released, the latch and the data it protects become available to other transactions. Since the modifications made by an epoch are buffered until it commits, we must postpone all release operations until after the epoch has fully committed. Release operations can be postponed by building a list of pending release operations as the epoch executes, and then performing all of the releases in the pending list when the epoch commits. If the epoch is violated, we simply reset this list.

Next we consider acquire operations. During normal execution, when a latch is acquired it prevents other transactions in the system from changing the



(a) Latch operations before re-ordering.



(b) Latch operations after re-ordering.

Figure 5: Delaying latch release operations until after a epoch commits can introduce deadlock.

associated data. A naïve approach to handling a latch acquire under TLS is to perform the acquire *non-speculatively* at the point when it is encountered. This can be implemented by a recursive latch, which counts the number of acquires and releases, and makes the latch available to other transactions only when the count reaches zero. This *aggressive* approach, shown in Figure 4(b), has a major drawback: since latch releases have been delayed until the end of the epoch, we have increased the overall size of the critical section. In addition, since we have parallel overlap between multiple critical sections in a single transaction, the latch may be held for an extended period of time.

To avoid long critical sections, we can also postpone acquires as shown in Figure 4(c). In this *lazy* approach, all latch acquires are performed at the end of the epoch, then the buffered speculative modifications are committed, and finally all latch releases are performed. This method results in much smaller critical sections, even when acquire and release operations for a given latch are encountered repeatedly during an epoch. A potential disadvantage of this approach is that if another transaction changes the protected data, the epoch will violate and restart.

Both the lazy and aggressive latch schemes have a potential problem: they re-order the latch release operations relative to the latch acquire operations as specified in the original program. If multiple latches are acquired by a single epoch, a deadlock may emerge that is not possible in the sequential execution, as shown in Figure 5. Although such deadlocks should be rare, there are two strategies to remedy them: *avoidance* and *recovery*. Deadlock can be avoided using two traditional techniques: (i) perform all latch acquires in a single atomic operation, or (ii) enforce a global latch

acquire ordering [17], such as by sorting the acquire queue by latch address. If avoidance is not possible, we can instead recover from deadlock once detected (perhaps through a time-out) by violating and restarting one of the deadlocked epochs. Forward progress is guaranteed because there is always at least one epoch (the oldest) which executes non-speculatively. The key insight is that restarting an epoch is much cheaper than restarting the entire transaction since there are many epochs per transaction.

4.1.2 Locks

Locks are a more sophisticated form of concurrency control than latches. Instead of providing simple mutual exclusion, locks allow multiple threads into a critical section at the same time if the lock types are *compatible*: multiple readers are allowed into a critical section at a time, while writers have exclusive access. Locks also provide deadlock detection, since multiple locks can be held at once and they are meant to be held for longer periods of time than latches.

We start by parallelizing locks using a *lazy locking* scheme, similar to the *lazy latch* scheme in Section 4.1.1. When an acquire operation is encountered in speculative code, we cannot simply delay the entire acquire operation until the end of the epoch, since a handle must be returned. Instead, we return an *indirect* handle, which is a pointer to an empty handle that is filled in at the end of the epoch when the lock acquire is actually performed.

To summarize our scheme so far, at the end of an epoch all of the lock acquires encountered in that epoch are performed, the changes made by the epoch are committed, and then all of the lock releases encountered in the epoch are performed. This scheme will result in correct execution, but holding all of the locks used by an epoch *simultaneously* can be a performance bottleneck in the database, particularly for the locks used for searching B-trees. We avoid this problem through a minor change: at the end of the epoch we (i) acquire *and release* all *read-only* locks in the order that the acquire and release operations were encountered during the epoch, we then (ii) perform all *non-read-only* lock acquires that were encountered during the epoch, (iii) commit the epoch's changes to memory, and then (iv) perform all *non-read-only* lock releases that were encountered during the epoch. Since a B-tree search involves briefly acquiring a large number of read-only locks, this ensures that those locks are held for minimal time; we need not hold the *read-only* locks during the epoch commit because the system view of an epoch commit is similar to a transaction commit: it either succeeds or fails. By acquiring and releasing the locks we ensure that the epoch commit does not occur in the middle of a non-read-only critical section in some other transaction.¹ If latches

¹Our method of executing lock acquires may also possibly

were labeled as read-only or read/write then this optimization could also be applied to latches in addition to locks.

4.1.3 Cursor Management

Cursors are data structures used to index into and traverse B-trees. Since they are used quite frequently and their creation is expensive, they are maintained in pre-allocated stacks. Unused cursors are stored in a *free cursor stack*. A dependence between epochs is created when one epoch puts a cursor onto the free cursor stack and the next epoch removes that cursor from the stack, since both operations manipulate the free pointer. Preserving this dependence is not required for correct execution: the second epoch did not need to get the exact same cursor, but instead wanted to get *any* cursor from the free stack. We can eliminate this dependence by partitioning the stack, and hence maintaining a separate stack for each processor. This implies that more cursors will have to be allocated, but that each cursor will only be used by the CPU which allocated it, increasing cache locality and eliminating dependences between epochs.

4.1.4 Memory Allocation

The free cursor pool mentioned above is just a special case of memory allocation. The general purpose memory allocators (such as `db_malloc`) in the database system introduce dependences between epochs when they update their internal data structures. To avoid these dependences, we must substitute an allocator designed with TLS in mind: in the common case, such an allocator should not communicate between CPUs. Fortunately, this is also a requirement of highly scalable parallel applications. The Hoard memory allocator [3] is one such allocator, which maintains separate free lists for each CPU, so that most requests for memory do not communicate. In the next section we show an even simpler way of avoiding dependences which does not require modifying or changing the underlying memory allocator.

4.1.5 Buffer Pool Management

When either a transaction or the DBMS itself need to read a page of the database, they request that page by invoking the `pin_page` operation on the buffer pool. This operation reads the requested page into memory (if it is not already there), pins it in memory, and returns a pointer to it. Once finished with the page, it is released by the `unpin_page` operation.

Conceptually, the buffer pool is very similar to the memory allocator, since it manages memory. However, the buffer pool is different because users explicitly

cause a deadlock situation. Similarly to latches, we can recover from a detected deadlock situation by violating and restarting one of the deadlocked epochs.

```

page_t *pin_page_wrapper(pageid_t id) {
    static intra_transaction_latch mut;
    page_t *ret;

    suspend_speculation();           // (iii)
    check_pin_page_arguments(id);    // (i)
    acquire_latch(&mut);             // (ii)

    ret = pin_page(id);

    release_latch(&mut);             // (ii)
    on_violation_call(unpin_page, ret); // (iv)
    resume_speculation();           // (iii)

    return ret;
}

```

Figure 6: Wrapper for the `pin_page` function which allows the ordering between epochs to be relaxed.

name the memory they want, and different `pin_page` operations can pin the *same page*. Therefore, simply partitioning the page pool between epochs will not suffice. Instead, we exploit the fact that the order in which `pin_page` operations take place *does not matter*. If a speculative epoch fetches the wrong page from disk, we simply must return that page to the free pool. We implement this by executing the `pin_page` function non-speculatively, so that it really does get the page and pin it in a way which is visible to the entire system. If the epoch which called `pin_page` is later violated, we can undo this action by calling `unpin_page`. (This is similar to the compensating transactions used in Sagas [4].)

The code wrapper shown in Figure 6 implements this modified version of `pin_page`. In particular, this code does the following: (i) Provides thorough error checking. Since this routine is called from a speculative thread, the parameters could be invalid. (ii) Acquires a latch which provides mutual exclusion between epochs within a transaction, to guard against the possibility that `pin_page` was not implemented with intra-transaction concurrency in mind. However, this latch can be eliminated if the programmer determines that the implementation of `pin_page` is safe in this respect. (iii) Temporarily suspends speculation. While speculation is suspended, the epoch is *non-speculative* and hence all reads will observe committed machine state and all writes will be immediately visible to the rest of the system (i.e., no buffering occurs). Hence reads performed by `pin_page` will not cause violations. (iv) Saves a pointer to the recovery function, `unpin_page`. If the epoch is violated then `unpin_page` will be called to undo the execution of the corresponding call to `pin_page`.

Relaxing ordering constraints simplifies coding: instead of redesigning the buffer pool to be amenable to TLS execution, we place this simple wrapper around

the allocation function. However, this method requires that the `pin_page` function be an *isolated undoable operation*. The `pin_page` function is undoable: calling `unpin_page` undoes the call to `pin_page`. The `pin_page` is also isolated: when it is undone via `unpin_page` no other transaction or earlier epoch is forced to rewind or otherwise alter its execution.

Similar reasoning shows that the cursor allocation function and `db_malloc` are also isolated undoable operations, and so this code template could be applied to these functions instead of partitioning their free pools. The `lock_acquire` and `latch_acquire` functions also look like isolated undoable operations—but as we found above in Section 4.1.1, without great care speculatively executing these functions out of original sequential order can cause performance problems (by increasing critical section sizes) or create deadlock conditions (by re-ordering lock and latch acquires).

The `unpin_page` operation for the buffer pool is *not* undoable, since an attempt to undo it with a `pin_page` operation may cause the page to be mapped at a different address. Because of this, we treat it similarly to a lock or latch release operation, and enqueue it to be executed after the epoch commits.

4.2 The Log

Every time the database is modified the changes are appended to the log. For recovery to work properly (using ARIES [12]) each log entry must have a log sequence number. Unfortunately, incrementing the log sequence number causes a data dependence between epochs. To avoid this dependence, we modify the logging code to append log entries for speculative epochs to a per-CPU buffer. When an epoch commits, we loop over this buffer to assign log sequence numbers to log entries, then append the entire buffer to the log.

4.3 B-Trees

B-trees are used extensively in the database system to index the database. The primary operations involving the B-tree are reading records, updating existing records, and inserting new records. Neither reading nor updating records modify the B-tree, and hence will not cause dependences between epochs. In contrast, insert operations modify the leaf pages of the B-tree. Therefore if the changes made by two epochs happen to fall on the same page then the update of the free space count for that page can cause a violation.

One strength of TLS parallelization is that infrequent data dependences need not be addressed, since the TLS mechanism will ensure correctness in such cases. An example of such an infrequent data dependence is a B-tree page split. Page splits can also cause many data dependences, but since they happen infrequently (by design), we can afford to just ignore them. In the rare cases when page splits occur, the TLS mechanism will ensure their correct sequential

execution. The TLS mechanism provides a valuable fall-back, allowing the programmer to avoid the effort of designing an algorithm for parallel page-splits.

The B-tree code in BerkeleyDB contains a simple performance optimization: when a search is requested, it begins the search by inspecting the page located by the previous search through a “last page referenced” pointer (this assumes some degree of locality in accesses). Accesses to this pointer cause a data dependence between epochs. Since the resulting violations can hurt performance, we decided to disable this “last page” optimization for TLS execution. Alternatively, one could retain this optimization without causing violations by maintaining a separate “last page reference” pointer per CPU.

4.4 Error Checks

Our study indicated that error checking code in the database system can occasionally cause dependences between epochs. The most important of these is a dependence caused by reference counting for cursors—a mechanism in the DBMS which tracks how many cursors are currently in use by a transaction, and ensures that none are in use when the transaction commits. Since this code is solely for debugging a transaction implementation, it can be safely removed once the transaction has been thoroughly tested.

4.5 False Sharing

To minimize overhead, the TLS mechanism tracks data dependences at the granularity of a cache-line. However, accesses to different variables which happen to be allocated on the same cache line can cause data dependence violations due to *false sharing*. This problem can be remedied by inserting padding to ensure that variables which are frequently-accessed by different CPUs are not allocated on the same cache line.²

5 Experimental Results

In this section we evaluate the ease with which a DBMS programmer can parallelize transactions, as well as the resulting performance.

5.1 Benchmark Infrastructure

We study the five transactions from TPC-C: (NEW ORDER, DELIVERY, STOCK LEVEL, PAYMENT and ORDER STATUS). We have parallelized both the inner and outer loop of the DELIVERY transaction, and denote the outer loop variant as DELIVERY OUTER. We also have modified the input to the NEW ORDER transaction to simulate a larger order of between 50 and 150 items (instead of the default 5 to 15 items),

²Insertion of padding works for most data structures, but is not appropriate for data structures which mirror disk-resident data, such as B-tree page headers. In this case, changes will have to be made to the B-tree data structure itself (see Section 4.3).

Table 1: Simulated memory system parameters.

Cache Line Size	32B
Instruction Cache	32KB, 4-way set-associative
Data Cache	32KB, 4-way set-associative, 2 banks
Unified Secondary Cache	2MB, 4-way set-associative, 4 banks
Speculative Victim Cache	64 entry
Miss Handlers	128 for data, 2 for insts
Crossbar Interconnect	8B per cycle per bank
Minimum Miss Latency to Secondary Cache	40 cycles
Minimum Miss Latency to Local Memory	75 cycles
Main Memory Bandwidth	1 access per 20 cycles

and denote that variant as NEW ORDER 150. All transactions are built on top of BerkeleyDB 4.1.25. Although a normal TPC-C run executes a concurrent mix of transactions and measures *throughput*, we are concerned with *latency*; hence we execute the individual transactions one at a time. Also, since we are primarily concerned with parallelism at the CPU level, we attempt to avoid I/O by configuring the DBMS with a large (100MB) buffer pool (this is roughly the size of the entire dataset for a single warehouse).

The parameters for each transaction are chosen using the Unix `random` function, and each experiment uses the same seed for repeatability. The benchmark executes as follows: (i) start the DBMS; (ii) execute 10 transactions to warm up the buffer pool; (iii) start timing; (iv) execute 100 transactions; (v) stop timing.

All code is compiled using gcc 2.95.3 with O3 optimization on a SGI MIPS-based machine. The BerkeleyDB database system is compiled as a shared library, which is linked with the benchmark that contains the transaction code.

To apply TLS to this benchmark we start with the unaltered transaction, mark the main loop within it as parallel, and execute it on a simulated system with TLS support. The system reports back the load and store program counters of the instructions which caused speculation to fail, and we use that information to determine the cause (in the source code) of the most critical performance bottleneck. We then apply the appropriate optimization from Section 4 and repeat.

5.2 Simulation Infrastructure

We perform our evaluation using a detailed, trace-driven simulation of a chip-multiprocessor composed of 4-way issue, out-of-order, superscalar processors similar to the MIPS R14000 [20], but modernized to have a 128-entry reorder buffer. Each processor has its own physically private data and instruction caches, connected to a unified second level cache by a crossbar switch. Register renaming, the reorder buffer, branch prediction (GShare with 16KB, 8 history bits), instruction fetching, branching penalties, and the memory hi-

erarchy (including bandwidth and contention) are all modeled, and are parameterized as shown in Table 1. Latencies due to disk accesses are not modeled, and hence these results are most readily applicable to situations where the database’s working set fits into main memory.

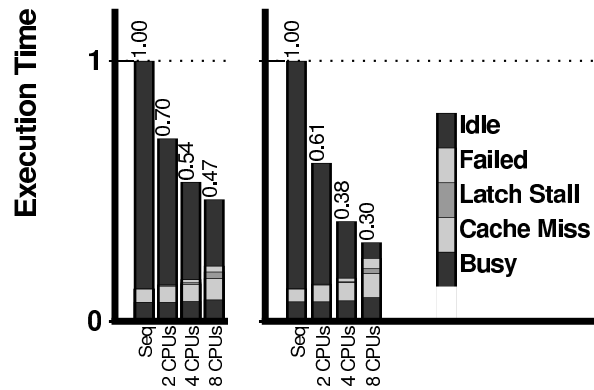
5.3 Scaling Intra-Transaction Parallelism

In Figure 7 we see the performance of the optimized transactions as the number of CPUs is varied. The SEQ bar represents the unmodified benchmark running on a single core of an 8 core chip multiprocessor, while the 2 CPU, 4 CPU and 8 CPU bars represent the execution of full TLS-optimized executables running on 2, 4 and 8 CPUs. Large improvements in transaction latency can be obtained by using 2 or 4 CPUs, although the additional benefits of using 8 CPUs are small. The PAYMENT and ORDER STATUS transactions are not shown because they do not benefit from TLS: PAYMENT contains no loops worth parallelizing, and the main loop in ORDER STATUS contains an unavoidable dependence.

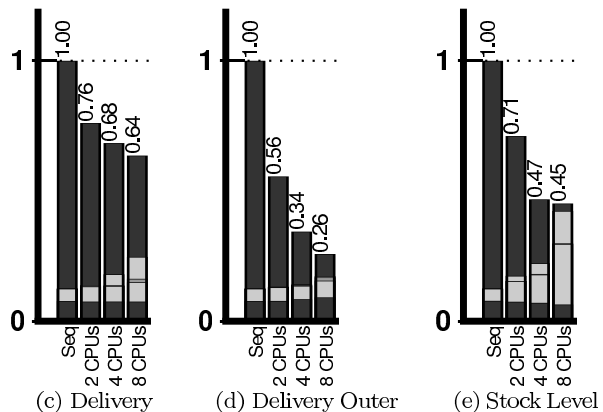
To better understand this data we break down each bar by where time is being spent. In Figure 7 we have normalized all bars to the 8 CPU case so that the subdivisions of each bar can be directly compared. This means that the SEQ breakdown shows one CPU executing and 7 CPUs idling, the 2 CPU breakdown shows two CPUs executing and 6 CPUs idling, etc.

The NEW ORDER, NEW ORDER 150 and DELIVERY OUTER bars show that very little time was spent on failed speculation—this means that our performance tuning was successful at eliminating performance-critical data dependences for those transactions. The DELIVERY transaction has some failed speculation due to a dependence in updating the ORDER LINE table, and the STOCK LEVEL transaction has failed speculation due to a dependence in the cursor used to scan the ORDER LINE table. As the number of CPUs increases there is a nominal increase in both failed speculation and time spent awaiting the latch used to serialize isolated undoable operations: as more epochs are executed concurrently, contention increases for both shared data and the latch. As the number of CPUs increases there is also an increase in time spent awaiting cache misses: spreading the execution of the transaction over more CPUs decreases cache locality, since the execution is partitioned over more level 1 caches. We also see a much larger increase in the number of cache misses for the STOCK LEVEL transaction—a large amount of cache state can be invalidated when speculation fails, leading to increased cache misses.

The dominant component of the bars in NEW ORDER and DELIVERY is *idle* time, for three reasons. First, in the SEQ, 2 CPU and 4 CPU case we show the unused CPUs as idle. Second, the loop that we paral-



(a) New Order (b) New Order 150



(c) Delivery (d) Delivery Outer (e) Stock Level

Category	Explanation
Idle	Not enough threads were available to keep the CPUs busy.
Failed	CPU executed code which was later undone due to a violation (includes all time spent executing failed code.)
Latch Stall	Stalled awaiting latch; latch is used in isolated undoable operations.
Cache Miss	Stalled on a cache miss.
Busy	CPU was busy executing code.

Figure 7: Performance of optimized benchmark while varying the number of CPUs.

lelized in the transaction only covers 78% of the transaction’s execution time for NEW ORDER, and 63% for DELIVERY: during the remaining time only one CPU is in use. Third, TPC-C specifies that both transactions will deal with orders which contain between 5 and 15 items, which means that on average each transaction will have only 10 epochs—this means that as we execute the last epochs in the loop load imbalance will leave CPUs idling. The effects of all three of these issues are magnified as more CPUs are added. To see the impact of reducing this idle time, we modified the invocation of the NEW ORDER transaction so that

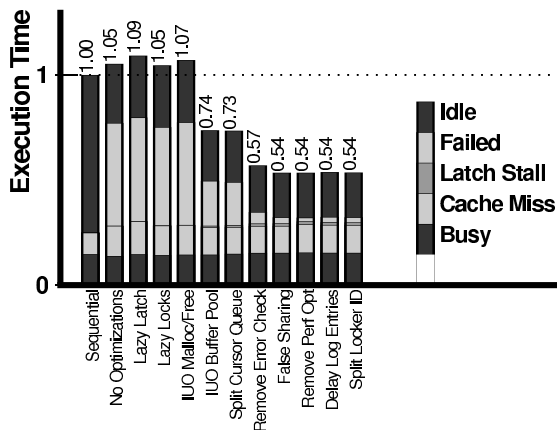


Figure 8: Performance impact on NEW ORDER of adding each optimization one-by-one on a four CPU machine.

each order contains between 50 and 150 items (which is the NEW ORDER 150 transaction). We found that this modification decreases the amount of time spent idling, and does not significantly affect the trends in cache usage, failed speculation, or idle time.

Figure 7 shows a performance trade-off: devoting more CPUs to executing a single transaction improves performance, but results in increased contention, a decrease in cache locality, and/or diminishing returns due to a lack of available parallelism, thus resulting in diminishing returns as more CPUs are added. One of the strengths of using TLS for intra-transaction parallelism is that it can be enabled or disabled at any time, and the number of CPUs can be dynamically tuned. The database system’s scheduler can dynamically increase the number of CPUs available to a transaction if CPUs are idling, or to speed up a transaction which holds heavily contended locks. If many epochs are being violated, and thus the intra-transaction parallelism is providing little performance benefit, then the scheduler could reduce the number of CPUs available to the transaction. If the transaction compiler simply emitted a TLS parallel version of *all* loops in transactions then the scheduler could use sampling to choose loops to parallelize: the scheduler could periodically enable TLS for loops which are not already running in parallel, and periodically disable TLS for loops which are running in parallel. If the change improves performance then the scheduler can make it permanent.

5.4 Impact of Each Optimization

In Figure 8 we see the results of the optimization process for the NEW ORDER benchmark on a four CPU system (the other transactions show similar trends). In this case the breakdown of the bars is normalized to a four CPU system, and so $\frac{3}{4}$ of the SEQUENTIAL bar is *Idle*, since three of the four CPUs are idling during the entire execution. The NO OPTIMIZATIONS

bar shows what happens if we parallelize the loop and make no other optimizations—the existing data dependences in the DBMS prevent any parallelism from being exploited, and the fact that we have taken a sequential transaction and run it on four CPUs has reduced cache locality, causing it to slow down slightly.

The major source of failed speculation in our newly-parallelized transaction are the reads and writes to latches; hence we perform the lazy latch optimization described in Section 4.1.1. This optimization fixes the first performance bottleneck, and exposes the next bottleneck which is in the lock code. The first optimization also results in a slight slowdown, since the next bottleneck merely delays detection of failed speculation (as illustrated in Figure 3(a))—hence more execution has to be rewound.

Once we have eliminated latches as a bottleneck, the next bottleneck exposed is in the locking subsystem. We remove the lock bottleneck by implementing lazy locks from Section 4.1.2. We continue to remove the bottlenecks one by one: applying the code template from Figure 6 to `db_malloc` and the `pin_page` operation, parallelizing the free cursor pool, removing dependence causing error checks (Section 4.4), adding padding to avoid violations due to false sharing (Section 4.5), removing the “last page referenced” pointer from the B-tree search code (Section 4.3), delaying the generation of log entries until epochs are ready to commit (Section 4.2), and parallelizing the assignment of locker ids.

It is tempting to look at Figure 8 and conclude that the most important optimization was parallelizing the buffer pool, since adding this optimization caused the execution time to drop by 40%. However, this is not the case since the impact of the optimizations is *cumulative*. If we take the NO OPTIMIZATIONS build and just enable the buffer pool optimization then the normalized performance is 0.98. Instead, Figure 8 implies that the iterative optimization process which we used works well—as the DBMS programmer removes performance limiting dependences performance gradually improves (and exposes new dependences). Removing dependences decreases the time spent on failed execution, and improves performance.

6 Conclusions

Chip multiprocessing has arrived, as evidenced by recent products (and announced road maps) from Intel, AMD, IBM and Sun Microsystems. While the database community has long embraced parallel processing, the fact that an application *must* exploit parallel threads to tap the performance potential of these additional CPU cores presents a major challenge for desktop applications. Processor architects have responded to this challenge through a new mechanism—*thread-level speculation* (TLS)—that enables optimistic parallelization on chip multiproces-

sors. Fortunately for the database community, although TLS was originally designed to overcome the daunting challenge of parallelizing desktop applications, it also allows us to tap new forms of parallelism within a DBMS that had previously been too painful to consider.

In this paper, we have focused on one such opportunity enabled by TLS: exploiting *intra-transaction* parallelism. Our experimental results demonstrate that we can speed up the *latency* (not just the throughput) of three of the five transactions in TPC-C by 36–74% by exploiting TLS on a chip multiprocessor with four CPU cores, or by 29–44% with two cores. TLS allows the database’s scheduler to use CPU cores to improve latency when throughput is not the primary concern. In contrast with previous approaches to exploiting intra-transaction parallelism, we place almost no burden on the transaction programmer (they merely demarcate epoch boundaries). In the future this burden could easily be shifted to the transaction compiler. Although changes to the DBMS code are required to achieve this benefit, they affected less than 1200 out of 180,000 lines of code in BerkeleyDB, they were implemented in roughly a month by a graduate student, and we expect that they would generalize to other DBMSs. We hope that these promising results will inspire database researchers to find other opportunities for exploiting untapped parallelism through TLS.

References

- [1] C.B. Colohan, A. Ailamaki, J.G. Steffan, and T.C. Mowry. Extending Thread Level Speculation Hardware Support to Large Epochs: Databases and Beyond. Technical Report CMU-CS-05-109, School of Computer Science, Carnegie Mellon University, March 2005.
- [2] IBM Corporation. *IBM DB2 Universal Database Administration Guide: Performance*. IBM Corporation, 2004.
- [3] E.D. Berger and K.S. McKinley and R.D. Blumofe and P.R. Wilson. Hoard: A Scalable Memory Allocator for Multithreaded Applications. In *Proceedings of the 9th ASPLOS*, 2000.
- [4] H. Garcia-Molina and K. Salem. Sagas. In *Proceedings of the 1987 ACM SIGMOD international conference on Management of data*, pages 249–259. ACM Press, 1987.
- [5] S. Gopal, T. Vijaykumar, J. Smith, and G. Sohi. Speculative Versioning Cache. In *Proceedings of the 4th HPCA*, February 1998.
- [6] J. Gray. *The Benchmark Handbook for Transaction Processing Systems*. Morgan-Kaufmann Publishers, Inc., 1993.
- [7] L. Hammond, B. Hubbert, M. Siu, M. Prabhu, M. Chen, and K. Olukotun. The Stanford Hydra CMP. *IEEE Micro Magazine*, March–April 2000.
- [8] M. Herlihy and J. Moss. Transactional memory: Architectural support for lock-free data structures. In *Proceedings of the 20th ISCA*, 1993.
- [9] H. Kaufmann and H.J. Schek. Extending tp-monitors for intra-transaction parallelism. In *Proceedings of the 4th PDIS*, 1996.
- [10] H.T. Kung and J.T. Robinson. On optimistic methods for concurrency control. *ACM TODS*, pages 213–226, June 1981.
- [11] J.H. Miller and H. Lau. *Microsoft SQL Server 2000 Resource Kit*, chapter RDBMS Performance Tuning Guide for Data Warehousing, pages 575–653. Microsoft Press, 2001.
- [12] C. Mohan, D. Haderle, B. Lindsay, H. Pirahesh, and P. Schwarz. ARIES: A Transaction Recovery Method Supporting Fine-Granularity Locking and Partial Rollbacks Using Write-Ahead Logging. *ACM TODS*, March 1992.
- [13] Michael Olson, Keith Bostic, and Margo Seltzer. Berkeley db. In *Proceedings of the Summer Usenix Technical Conference*, June 1999.
- [14] M.K. Prabhu and K. Olukotun. Using thread-level speculation to simplify manual parallelization. In *The ACM SIGPLAN 2003 Symposium on Principles & Practice of Parallel Programming*, June 2003.
- [15] M. Rys, M.C. Norrie, and H.J. Schek. Intra-transaction parallelism in the mapping of an object model to a relational multi-processor system. In *Proceedings of the 22nd VLDB*, 1996.
- [16] D. Shasha, F. Llirbat, E. Simon, and P. Valduriez. Transaction chopping: Algorithms and performance studies. *ACM TODS*, 20(3):325–363, 1995.
- [17] A. Silberschatz, P.B. Galvin, and G. Gagne. *Operating System Concepts*. John Wiley & Sons, Inc., 2002.
- [18] J.G. Steffan, C.B. Colohan, A. Zhai, and T.C. Mowry. A Scalable Approach to Thread-Level Speculation. In *Proceedings of the 27th ISCA*, June 2000.
- [19] M. Tremblay. MAJC: Microprocessor Architecture for Java Computing. *HotChips '99*, August 1999.
- [20] K. Yeager. The MIPS R10000 superscalar microprocessor. *IEEE Micro*, April 1996.
- [21] C. Zuzarte. Personal communication, 2005.