Capriccio: Scalable Threads for Internet Services

F Z u G . N cula a E c w
pu Sc c D s
U s f alf a k l
{j jc zf cula w }@cs. k l . u

ABST ACT

T is paper prese ts apri i a s a a e t rea pa a e r se it i rre ser ers. ie re e t r as a ate e e t ase s stems e e ie e t att rea ase s stems a pr i e a simp er pr rammi m e t at a ie es eq i a e t r s peri r per rma e.

B imp eme ti apri i as a ser e e t rea pa a e e a e e p e t e t rea pa a e imp eme tati r m t e er i perati s stem. As a res t e a ta e a a ta e perati e t rea i e as r s I/O me a isms a mpi er s pp rt. Usi t is appr a e are a e t pr i e t ree e eat res: (1) s a a i it t 100 000 t rea s (2) effi ie t sta ma a eme t a (3) res r e a are s e i .

e i tr e *linked stack anage ent* i mi imizes t e am t aste sta spa e pr i i sa e sma a ti s sta s t at a r r s ri at r time. A mpi er a a sis ma es r sta imp eme tati effi ie t a s . e as prese t *resource-a are sc edul-ing* i a st rea s e i a a missi tr t a apt t t es stem s rre t res r e sa e. T is te iq e ses a *locking gra* t at is a t mati a eri e r m t e app i ati t es ri e t e tr et ee i p its i a perati e t rea pa a e. e a e app ie r te iq es t t e Apa e 2.0. e ser er em strati t at e a a ie e i per rma e a s a a i it espite si a simp e t rea e pr rammi m e.

C S j

. .1 : r ess a a eme t *t reads*

T

A rit ms esi er rma e

ser e e t rea s i e sta ma a eme t ami sta r t res r e a are s e i i rap

SOSP'03, October 19–22, 2003, Bolton Landing, New York, USA.

T CT

T a s I ter et ser i es a e e er i reasi s a a i it ema s. er ser ers m st e apa e a i te s r re s t sa s sim ta e s e ti s it t si i a t per rma e e ra ati . rre t m m it ar are is apa e meeti t ese ema s t s t are as a e e i . I parti ar t ere is a pressi ee r a pr rammi m e t at a s pr rammers t esi effi ie t a r st ser ers it ease.

T rea pa a es pr i e a at ra a stra ti r i rre pr rammi t i re e t ears t e a e ee s pp a te e e t ase s stems s as A 3 . T ese e e t ase s stems a e req ests si a pipe i e sta es. a req est is represe te a e e t a ea sta e is imp eme te as a e e t a er. T ese s stems a pre ise tr er at pr essi state ma a eme t a a missi tr i a iti t e pr i e e e ts s as at mi it it i ea e e t a er.

U rt ate e e t ase pr rammi as a m er ra a s e mpare t t rea e pr rammi 3 . e ts stems i e t e tr t r a app i ati ma i it iffi t t ersta a se a e e t re ati s ips e e ami i s r e ea e e e i . r i sta e ma e e t s stems i e a met i a t er m e se i a a e e t a t e aiti r a ret r e e t i resp se. I r er t ersta t e app i ati t e pr rammer m st me ta mat t ese a /ret r pairs e e e t e are i i ere t parts t e e. rt erm re reati t ese a /ret r pairs te req ires t e pr rammer t ma a sa e a rest re i e state. T is pr ess re erre t as sta rippi 1 is a ma r r e r pr rammers is t se e e t s stems.

I t is paper

I ee r t rea pa a e a impr e per rma e e isti t rea e appi ati s it itt e t m i ati t e appi ati itse .

Th P

I t e pr ess i t rea s r se i ser er appi a ti s e t ata ser e e appr a isesse tia. i e ser e e t rea s a er e t rea s are t se t e s e ame ta i ere t pr ems. Ker e t rea s are primari se r e a i tr e rre ia m tipe e ies is req ests r Us. User e e t rea s are rea i a t rea s t at s pr i e a ea pr rammi m e it se i aria ts a sema ti s.

T ate e t str a ate a *articular* se ma ti s r t rea s rat er e ar e t at a ea se ma ti s r t rea s req ires *decou ling* t e t rea s t e pr rammi m e (i a t rea s) r m t se t e er i er e. e pi t e pr rammi m e r m t e er e is imp rta t r t reas s. irst t ere is s sta tia ariati i i ter a es a sema ti s am m er er e s espite t e e iste e t e O IX sta ar . e er e t rea s a as r s I/O i ter a es are areas a ti e resear 19 20. T e ra e sema ti s a t e rate e ti t req ire e pi : i a t rea s a i e t O ariati a er e e ti .

I r ase t is e pi as pr i e a m er a a ta es. e a e ee a e t i te rate mpi er s pp rt i t r t rea pa a e a e a e ta e a a ta e se era e er e eat res. T s e a e ee a e t i rease per rma e impr e s a a i it a a ress appi ati spe i ee s a it t a i appi ati e.

2 C

T is paper is sses r e t rea pa a e apri i . T ist rea pa a e a ie es r as it t e e p t ree e eat res:

irst e impr e t e s a a i it asi t rea pera ti s. e a mp is e t is tas si ser e e t rea s it perati e s e i ta i a a ta e a e as r s I/O i ter a e a e i eeri r r time s stem s t at a t rea perati s are $O(1)$.

e e i tr e *linked stacks* a me a ism r ami sta r t t at s es t e pr em sta a ati r ar e m ers t rea s. Tra iti a t rea s s tems prea ate ar e s mem r rea t rea s sta i se ere imits s a a i it. apri i ses a m i ati mpi e time a a sis a r time e s t imit t e am t aste sta spa e i a effi ie t a appi ati spe i ma er.

i a e esi e a *resource-a are sc eduler* i e tra ts i rmati a t t e tr it i a pr ram i r er t ma e s e i e isi s ase pre i te res r e sa e. T is s e i te iq e ta es a a ta e mpi er s pp rt a perati e t rea i t a ress appi ati spe i ee s it t req iri t e pr rammer t m i t e ri i a pr ram.

T e remai er t is paper is sses ea t ese t ree eat res i etai. T e e prese t a era e perime ta e a ati r t rea pa a e. i a e is ss t re ire ti s r ser e e t rea pa a es it i te rate mpi er s pp rt.

2 TH EA ES A SCALAB L TY

apri i is a ast ser e e t rea pa a e t at s pp rts t e O IX A I r t rea ma a eme t a s r iza ti . I t is se ti e is ss t e era esi r t rea pa a e a e em strate t at it satis es r s a a i it as.

2 -L v Th

O e t e rst iss es e e p re e esi i apri i as et er t emp ser e e t rea s r er e t rea s. User e e t rea s a es me imp rta t a a ta es r t per rma e a e i i it. U rt ate t e as m p i ate preempti a a i tera t a it t e er e s e er. U timate e e i e t at t e a a ta es ser e e t rea s are si i a t e t arra t t e a iti a e i eeri req ire t ir m e t t eir ra a s.

2.1.1 Flexibility

User e e t rea s pr i e a treme s am t e i i it r s stem esi ers reati a e e i ire ti et ee appi ati s a t e er e. T is a stra ti e ps t e pe t e t a it a s aster i ati t si es. r e amp e apri i is apa e ta i a a ta e t e e as r s I/O me a isms t e e e pme t series Li er e i a s s t pr i e per rma e impr eme ts it t a i appi ati e.

T e se ser e e t rea s as i reases t e e i i it t e t rea s e er. Ker e e e t rea s e i m st e e era e t pr i e a reas a e e e q a it r a appi ati s. T s er e t rea s a t tai r t e s e i a rit m t t a spe i appi ati . rt ate ser e e t rea s t s er r m t is imitati . I stea t e ser e e t rea s e er a e i ta it t e appi ati .

User e e t rea s are e treme i t ei t i a s pr rammers t se a treme s m er t rea s it t rr i a t t rea i er e a . T e e mar s i e ti 2.3 s t at apri i a s a e t 100,000 t rea s t s apri i ma es it p ssi e t rite i rre t appi ati s (i are te ritte it mess e e t ri e e) i a simp e t rea e st e.

2.1.2 Performance

User e e t rea s a reat re e t e er e a t rea s r izati . I t e simp est ase perati es e i a si e U s r izati is ear ree si e eit er ser t rea s r t e t rea s e er a e i ter r pte ie i a riti a se ti .[1] I t e t re e e ie e t at e i e ser e e s e i a mpi e time a a sis i a s t er simi ar a a ta es a m ti U ma i e.

e i t e ase preempti e t rea i ser e e t rea s er a a a ta e i t at t e t req ire er e r ss i s r m te a q isiti r re ease. B mparis er e e e m t a e si req ires a er e r ssi r e er s r izati perati . ie t is sit ati a e im pr e r te e s[2] i te e m te es sti req ire er e r ssi s.

[1] r esi e si a a i e a rei tr e t ese pr ems t t is pr em a easi e a i e .

[2] T e `futexes` i re e t Li er e s a perati s te e m te es t r e tire i ser spa e.

i a mem r ma a eme t is m re effi ie t it ser e e t rea s. Ker e t rea s req ire ata str t res t at eat p a a e er e a ress spa e e reasi t e spa e a aia e r I/O ers e es ript rs a t er res r es.

2.1.3 Disadvantages

User e e t rea i is t it t its ra a s e er. I r er t retai tr t e pr ess r e a ser e e t rea e e tes a i I/O a a ser e e t rea i pa a e erri es t ese i a s a rep a es t em i ter a it i eq i a e ts. T e sema ti s t ese i I/O me a isms e era req ire a i rease m er er e r ssi s e mpare t e i eq i a e ts. r e amp e t e m st effi ie t i et r I/O primiti e i Li (`epoll`) i es rst p i s ets r I/O rea i ess a t e per rmi t e a t a I/O a . T ese se I/O a s are i e ti a t t se per rme i t e i ase t e p a s are a iti a er ea . N i is I/O me a isms are te simi ar i t at t e emp separate s stem a s t s mit req ests a retrie e resp ses.[3]

I a iti ser e e t rea pa a es m st i tr e a rapper a er t at tra s ates i I/O me a isms t i I/O es a t is a er is a t ers r e er ea . At est t is a er a e a er t i s im i simp a s a e e tra ti a s. H e er r q i perati ss as i a e rea s t at are easi satis e t e er e t is er ea a e me imp rta t.

i a ser e e t rea i a ma e it m re iffi t t ta e a ta a e m tip e pr ess rs. T e per rma e a a ta a e i t ei ts r izati is imi is e e m tip e pr ess rs are a e si e s r izati is er r ree . A iti a as is sse A ers et a . i t eir r s e er a ti ati s p re ser e e s r izati me a isms are i e e ti e i t e a e tr e rre a ma ea t star ati 2 .

U timate e e ie e t e e e ts ser e e t rea i ar t ei t ese isa a ta es. As t e e mar s i e ti 2.3 s t e a iti a er ea i rre es t seem t e a pr em i pra ti e. I a iti e are r i a s t er me t e iffi ties it m tip e pr ess rs e i is ss t is iss e rt er i e ti 7.

2 2

e a e imp eme te apri i as a ser e e t rea i i rar r Li . apri i imp eme ts t e O IX t rea i A I i a s it t r m st appi ati s it t m i ati .

Co x w ch . apri i is i t t p ar T er i s r ti e i rar 32. T is i rar pr i es e treme ast te t s it es r t e mm ase i i t rea s tari ie eit er e p i it r t r ma i a i I/O a . e are rre t esi i si a ase e t at a s r preempti r i ser t rea s t apri i es t pr i e t is eat re et.

I/ . apri i i ter epts i I/O a s att e i rar e e erri i t e s stem a st ti s i GNU i . T is appr a r s a ess r stati a i e appi ati s a r ami a i e appi ati s t at se GNU i ersi s 2.2 a ear ier. H e er GNU i ersi 2.3 passes t e s stem a st s r ma its i ter a r ti es (s as `printf`) i a ses pr ems r ami a i e appi ati s. e are r i t a apri i t ti as a i a i r er t pr i e etter i te rati it t e atest ersi s GNU i .

I ter a apri i ses t e atest Li as r s I/O me a isms `epoll` r p a e e es ript rs (e. . s ets pipes a s) a Li AIO r is . I t ese me a isms are t a aia e apri i a s a t e sta ar U i `poll()` a r p a e es ript rs a a p er e t rea s r is I/O. Users a se e t am t e a aia e I/O me a isms setti appr priate e i r me t aria es pri r t starti t eir appi ati .

ch dul . apri i s mai s e i p s er m i e a e e t ri e appi ati a ter ate r i appi ati t rea s a e i r e I/O mp eti s. N te t t at t e s e er i es t is e e t ri e e a i r r m t e pr rammer sti ses t e sta ar t rea ase a stra ti . apri i as a m ar s e i me a ism t at a s t e ser t easi se e t et ee i ere t s e ers at r time. T is appr a as as ma e it simp e r st e e p se era i ere t s e ers i i a e s e er ase t rea res r e ti iza ti . e is ss t is eat re i etai i e ti .

ch o z o . apri i ta es a a ta e per ati e s e i t impr e s r izati . At prese t apri i s pp rts perati e t rea i si e U ma i es i i ase i ter t rea s r izati primiti es req ire simp e e s a ea e / e a . r ases i i m tip e er e t rea s are i e apri i emp s eit er spi s r ptimisti r re tr primiti es epe i i me a ism est ts t e sit ati .

Effic c . I e e pi apri i e a e ta e reat are t se effi ie t a rit ms a ata str t res. seq e t a t e apri i s t rea ma a eme t ti s as a e rst ase r i time i epe e t t e m er t rea s. T e s e e epti is t e seep q e e i rre t ses a ai e i e ist impe me tati . i e t e iterat re tai s a m er a rit ms r effi ie t s eep q e es r rre t imp eme tati as t a se pr ems et s e a e se r e e pme t e rts t er aspe ts t e s stem.

2 3 Th M h k

e ra a m er mi r e mar s t a i ate apri i s esi a imp eme tati . O r test p at rm as a it t 2. GHz Xe pr ess rs 1 GB mem r t 10K R I U tra II ar ri es a 3 Gi a it t er et i ter a es. T e perati s stem as Li 2.5.70 i i es s pp rt r `epoll` as r s is I/O a i t ei ts stem a s (`vsyscall`). e ra r e mar s t ree t rea pa a es: apri i Li T rea s (t e sta ar Li er e t rea pa a e) a N TL ersi 0.53 (t e e Nati e O IX T rea s r Li pa a e). e i t a appi ati s it 3.3 a i e

[3] A t t ere are i I/O me a isms (s as O IX AIO s `lio_listio()` a Li s e `io_submit()`) t at a t e s missi m tip e I/O req ests it a si e s stem a t ere are t er iss es t at ma e t is eat re iffi t t se. r e amp e imp eme tati s O IX AIO te s er r m per rma e pr ems. A iti a se at i reates a tra e et ee s stem a er ea a I/O ate i is iffi t t ma a e.

	Capriccio	Capriccio_notrace	LinuxThreads	NPTL
Thread creation	21.5	21.5	37.9	17.7
Thread context switch	0.56	0.24	0.71	0.65
Uncontended mutex lock	0.04	0.04	0.14	0.15

T bl 1: L c of h d v fo d ff h d ck .

a ai st GNU i 2.3. e re mpi e Li T rea s t se t e e i t ei t s stem a eat re atest Li er e s t e s rea air mparis it N TL i ses t is eat re.

2 4 Th P v

Ta e 1 mpares a era e times se era t rea prim iti es r apri i Li T rea s a N TL. I t e test a e e `Capriccio_notrace` e isa e statisti s e ti a ami sta a tra i (se r t e s e er is sse i e ti) t s t eir impa t per rma e. T rea reati time is mi ate sta a ati time a is q ite e pe si e r a r t rea pa a es. T rea te t s it es e er are si i a t aster i apri i e e it t e sta tra i a statisti s e ti er ea . e e ie e t at re e er e r ssi s a r simp er s e i p i t tri te t t is res t. r izati primiti es are a s m aster i apri i (a a t r r te e m te i) e a se er e r ssi s are i e .

2 5 Th S

T meas re t e era effi ie a s a a i it s e i a s r izati i i ere t t rea pa a es e ra a simp e pr er s mer mi r e mar t e t ree pa a es. r ers p t empt messa es i t a s are er a s mers pr ess ea messa e pi r a ra m am t time. r izati is imp eme te si iti aria es a m te es. q a m ers pr ers a s mers are reate r ea test. Ass i i re 1 apri i tper rms N TL a Li T rea s i terms t ra per rma e a s a a i it. T r p t Li T rea s e i s t e ra e q i ater 20 t rea s are reate a N TL s t r p t e ra es a ter 100. apri i s a es t 32K pr ers a s mers (K t rea s t ta). e attri te t e r p t r p t et ee 100 t rea s a 1000 t i rease a e tpri t.

2 6 / P f

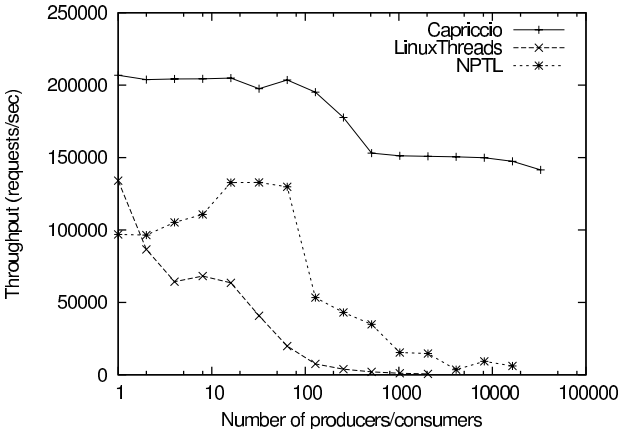
i re 2 s s et r per rma e apri i a t er t rea pa a es er a . I t is test e meas re t e t r p t rre t passi a m er t e s am a e m er pipes. T e m er r re t t e s is e q arter t e m er pipes i t ere are ess t a 12 pipes t er ise t ere are e a t 12 t e s. T e e mar t s sim ates t e e e t s ie t i s t at is a ar e m er m st i e pipes. T is s e ari is t pi a r I ter et ser ers a tra iti a t rea i s stems te per rm p r i s tests. T ti a eq i a e t e mar pr rams are se t tai t e res ts: a t rea e ersi is se r apri i Li T rea s a N TL a a i I/O ersi is se r `poll` a `epoll`.

T e re s s t at apri i s a es sm t t K t rea s a i rs ess t a 10% er ea e mpare t `epoll` it m re t a 25 pipes. T r e e `epoll` is t e est i I/O me a ism a ai a e Li e e its per rma e s re e t t at t e est e e t ase ser ers i a re s a me a ism. apri i per rms siste t etter t a Li T rea s a N TL it m re t a 25 t rea s a is m re t a t i e as ast as t Li T rea s a N TL e m re t a 1000 t rea s are reate .

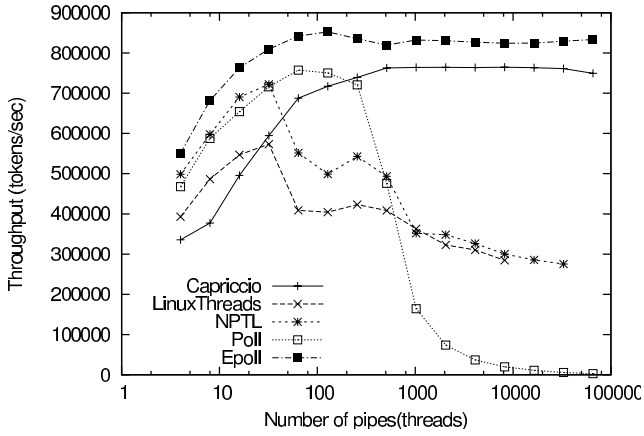
H e er e rre is (< 100 pipes) apri i is s er t a its mpetit rs e a se it iss es m re s stem a s. I parti ar it a s `epoll_wait()` t tai e es ript r rea i ess e e ts t a e p t rea s i r I/O. It per rms t ese a s peri i a tra s er ri as ma e e ts as p ssi e ea a . H e er e rre is t e m er r a e t rea s asi a rea es zer r i apri i t iss e m re `epoll_wait()` a s. I t e rst ase apri i is 37% s er t a N TL e t ere are 2 rre t t e s (a t rea s). rt ate t is er ea is am rtize q i e rre i reases m re s a a e s e i a s apri i t tper rm Li T rea s a N TL at i rre .

i e apri i ses as r s I/O primiti es apri i a e e t r m t e er e s is ea s e i a rit m st as m as er e t rea s a . i re 3 s s a mi r e mar i i a m er t rea s per rm ra m KB rea s r m a 1 GB e. T e test pr ram passes t e er e er a e si `O_DIRECT` e pe i t e e. T r p t a t ree t rea i raries i reases stea i it t e rre e e ti it e es e rre rea es a t 100. I trast ti izati t e er e s ea s e i a rit m i e e t ase s stems t at se i is I/O (e. . A) is imite t e m er er e t rea s se i is te ma e e i erate sma t re e er e s e i er ea . e rse t er pr ess ase app i ati st at se i I/O (eit er `poll()` `select()` `/dev/poll` r `epoll`) a t e e t r m t e er e s ea s e i at a i t e t e pi it se as r s I/O. U rt ate m st pr rams t se as r s I/O e a se it si i a t i reases pr rammi mp e it a mpr mises p rta i it .

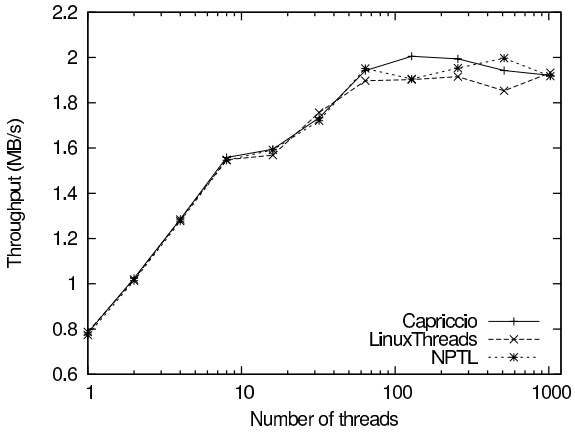
i re s s is I/O per rma e t e t ree t rea i raries e si t e O er a e. I t is test e meas re t e t r p t a ie e e 200 t rea s rea ti s r m t e e s stem it a spe i e er a e miss rate. T e a e miss rate is e rea i a appr priate p rti ata r m a sma e pe e rma (e e a a e its) a rea i t e remai i ata r m a e pe e it `O_DIRECT`. r a i er miss rate t e test is is t s apri i s per rma e is i e ti a t t at N TL a Li T rea s. H e er e t e miss rate is er t e pr ram is U s t r p t is imite per tra s er er ea . Here apri i s ma im m t r p t is a t 50% N TL s i mea s apri i s er ea is t i e t at N TL. T e s re t is er ea is t e as r s I/O i ter a e (Li AIO) se apri i i i rs same



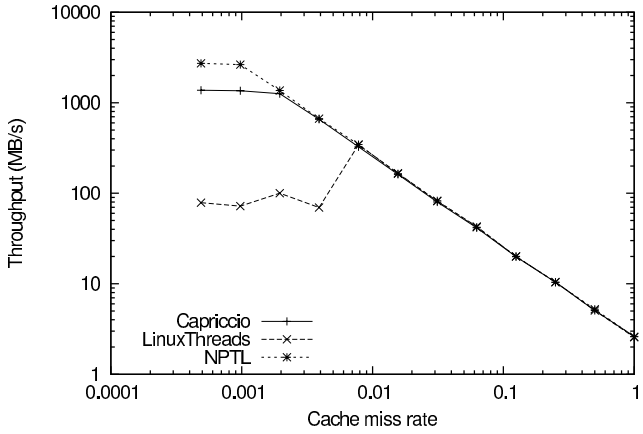
F u 1: Producer-Consumer - ch dul d ch o z o fo c .



F u 2: Pipetest - wo k c l b l .



F u 3: B fi of d k h d ch dul .



F u 4: D k I/ fo c w h buff c ch .

am t er ea r a e itti perati s a es t at rea es t e is : r ea I/O req est a mp eti e e t ee s t e str te q e e a e i ere t ser e e t r a separate s stem a . H e er t is s rt mi is reati e eas t : ret r i t e res t imme iate r req ests t at t ee t ait e a e imi ate m st (i t a) t is er ea . e ea e t is m i ati as t re r . i a a s rprisi res t is t at Li T rea s per rma e e ra es si i a t at a er miss rate. e e ie e t is e ra ati is a res t a eit er i t e er e r i t e i rar si e t e pr ess r is m st i e ri t e test.

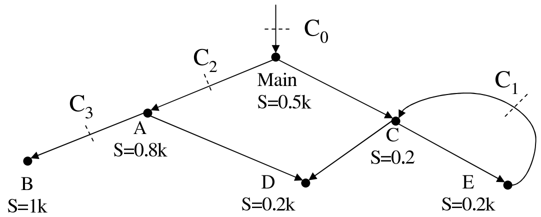
3 L E STAC MA A EME T

T rea pa a es s a attempt t pr i e t e pr ram mer it t e a stra ti a e a sta rea t rea . I rea it t e sta size is e t t e s are se ser ati e s t at t ere is p e t spa e r rma pr ram e e ti . r e amp e Li T rea s a ates t me a tes per sta e a t it s a ser ati e a ati s eme e s me 1 GB irt a mem r r sta spa e it st 500 t rea s. rt ate m st t rea s s me a e i tes sta spa e at a i e time at t e mi t t r sta es e t e se si era m re. T is ser ati s ests t at e a si i a t re e t e size irt a mem r e i ate t sta s i e a pt a ami sta a ati p i erei sta spa e is a ate t t rea s ema i reati e sma i reme ts a is ea ate e t e t rea req ires ess sta spa e. I t e rest t is se ti e is ss a mpi er eat re t at a s st pr i es a me a ism ie presen i t e pr rammi a stra ti e sta s.

3 C A L k S k

O r appr a ses a mpi er a a sis t imit t e am t sta spa e t at m st e prea ate . e per rm a e pr ram a a sis ase a *eig ted call gra* .[4] a ti i t e pr ram is represe te a e i t is a rap ei te t e ma im m am t sta spa e t at a si e sta rame r t at ti i s me. A e e et ee e *A* a e *B* i i ates t at ti *A* a s ti *B* ire t . T s pat s et ee es i

[4] e se t e IL t it 23 r t is p rp se i a s effi ie t e pr ram a a sis rea r appi ati s i e t e Apa e e ser er.



F u 5: A x l of c ll h o d w h ck f z . Th d k d w h C_i (=0,...,3) h ch ck o .

t is rap rresp t seq e es sta rames t at ma appear t e sta at r time. T e e t a pat is t e s m t e ei ts a es i t is pat t at is it is t e t ta size t e rresp i seq e e sta rames. A e amp e s a rap is s i i re 5.

Usi t is a rap e is t p a e a reas a e t e am t sta spa e t at i e s me ea t rea . I t ere are re rsi e ti s i r pr ram t ere i e es i t e a rap a t s e a easi t e ma im m sta size r t e pr ram at mpi e time i t e est pat starti r m ea t rea s e tr p i t. H e er m st rea r pr rams ma e se re rsi i mea s t at e a t mp te a t e sta size at mpi e time. A e e i t e a se e re rsi t e stati mp tati sta size mi t e t ser ati e. r e amp e si er t e a rap i i re 5. I ri t e e i t e rap t e ma im m sta size is 2.3 KB t e pat `Main–A–B`. H e er t e pat `Main–C–D` as a sma er sta size 0.9 KB. I t e rst pat is se ri i itia izati a t e se pat is se t r t e pr ram s e e ti t e a ati 2.3 KB t ea t rea e aste . r t ese reas s it is imp rta t t e a e t r a s ri t e sta size ema .

I r er t imp eme t ami a size sta s r a rap a a sis i e ti es a sites at i e m st i sert *c eck oints*. A e p i t is a sma pie e e t at etermi es et er t ere is e sta spa e e t t rea t e e t e p i t it t a si sta er . I t e spa e remai s a e *stack c unk* is a ate a t e sta p i ter is a ste t p i t t t is e . e t e ti a ret r s t e sta is i e a ret r e t a ree ist.

T is s eme res ts i ti s sta s t e a se t e sta s are s it e ri t e re t e a t a ar me ts r a ti a are p s e t e e r t e a ee ee t e a e . A e a se t e a ers rame p i ter is st re t e a ees sta rame e ers a t e a tra e a pr ram.[5] T e e r a e p i t is ritte i it a sma am t i i e assem r rea i a setti t e sta p i ter t is e is i serte si a s r e t s r e tra s rmati t e pr ram pri r t mpi ati . t a e si r a essi t e ree sta ist is e s re r perati e t rea i appr a .

[5]T is s eme es t r e t e `omit-frame-pointer` is e a e i `gcc`. It is p ssi e t s pp rt t is ptimizati si m re e pe si e e p i t perati s s as p i t e ar me ts r m t e a ers rame t t e a ees rame.

3 2 P ch k

ri r pr ram a a sis e m st etermi e ere t p a e e p i ts. A simp e s ti is t i sert e p i ts at e er a site e er t is appr a is pr i iti e e pe si e. A ess restri ti e appr a is t e s re t at at ea e p i t e a e a t e sta spa e t at ma e s me e re e rea t e e t e p i t (r a ea i t e a rap).

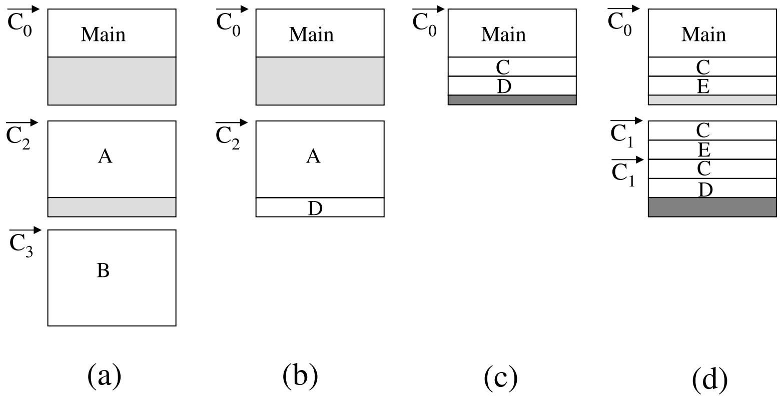
T satis t is req ireme t e m st e s re t at t ere is at east e e p i t i e er e it i t e a rap (re a t at t e e es i t e a rap rresp t a sites). T t e appr priate p i ts t i sert e p i ts e per rm a ept rst sear t e a rap i i e ti es a e es t at is e es t at e t a e t e its a est rs i t e a rap 22 . A es i t e rap m st tai a a e e s ea e p i ts at a a sites i e ti e as a e es i r ert t e s re t at a pat r m a ti t a e p i t as e e t . I i re 5 t e e p i t C_0 a ates t e rst sta a t e e p i t C_1 is i serte t e a e e `E–C`.

e a ter e rea a es t e s sta size ma e t ar e. T s e is t a a iti a e p i ts t t e rap i r ert t e s re t at a pat s et ee e p i ts are it i a esire i is i e as a mpi e time parameter. T i sert t ese e e p i ts e pr ess t e a rap e m re t is time etermi i t e est pat r m ea e t t e e t e p i t r ea . e per rmi t is a a sis e si er a restri te a rap t at es t tai a a e es si e t ese e es a rea a e e p i ts. T is restri te rap as es s e a pr ess t e es tt m p t s e pr essi e n e i a e a rea etermi e t e est pat r ea n s s ess rs. r ea s ess r s e n e ta e t e est pat r s a a n. I t is e pat s e t e ee s t e spe i e pat imit parameter e a a e p i t t t e e e et ee n a s i e e ti e re es t e est pat s t zer . T e res t t is a rit m is a set e es ere e p i ts s e a e a it reas a e s t e ma im m pat e t r m ea e. r t e e amp e i i re 5 it a imit 1 KB t is a rit m p a es t e a iti a e p i ts C_2 a C_3. it t t e e p i t C_2 t e sta rames `Main` a `A` se m re t a 1 KB.

i re s s r i sta es i t e i etime t e t rea se a rap is s i i re 5. I i re (a) t e ti B is e e ti it t ree sta s a ate at e p i ts C_0 C_2 a C_3. N ti e t at 0.5 KB is aste i t e rst sta a 0.2 KB is aste i t e se . I i re ( ) ti A as a e D a t sta s ere e essar . i a i i re ( ) e see a i sta e it re rsi . A e sta is a ate e E a s C (at e p i t C_1). H e er t e se time ar t e e at e p i t C_1 e i es t at t ere is e spa e remai i t e rre t sta t rea eit er a ea ti (D) r t e e t e p i t (C_1).

3 3 h S

ti p i ters prese t a a iti a a e e t r a rit m e a se e t at mpi e time e a t i ti ma e a e t r a i e ti p i ter. T impr e t e res ts r a a sis t e a t t etermi e as pre ise as p ssi e t e set



F u 6: Ex l of d c lloc o d d lloc o of ck chu k .

ti s t at mi t e a e at a ti p i ter a site. rre t e ate rize ti p i ters m er a t pe ar me ts t i t e t re e pa t se a m re s p isti ate p i ter a a sis.

a s t e ter a ti s a s a se pr ems si e it is m re iffi t t t e sta spa e se pre mpi e i raries. e pr i e t s ti s t t is pr em. irst e a t e pr rammer t a tate e ter a i rar ti s it tr ste sta s. A ter ati e e a ar er sta s t e i e r e ter a ti s as as t rea s t req e t it i t ese ti s e a re se a sma m er ar e sta s t r t t e app i ati . r t e sta ar i rar e se a tati s t ea it ti s t at r ti s t at are req e t a e t ese a tati s ere eri e a a zi i rar e.

3 4 T h A h

O r a rit m a ses sta spa e t e aste i t p a es. irst s me sta spa e is aste e a e sta is i e e a t is spa e *internal asted s ace.* e sta spa e at t e tt m t e rre t is si ere se t is spa e is a e *external asted s ace.* I i re i ter a aste spa e is s i i t ra ereas e ter a aste spa e is s i ar ra .

T e ser is a e t t e t parameters t at a st t e tra e s i terms aste spa e a e e ti spee . irst t e ser a a st *MaxPat* i spe i es t e ma im m esire pat e t i t e a rit m e a e st es ri e . T is parameter a e ts t e tra e et ee e e ti time a i ter a aste spa e ar er pat e t s req ire e er e p i ts t m re sta i i . e t e ser a a st *MinC unk* t e mi im m sta size. T is parameter a e ts t e tra e et ee sta i i a e ter a aste spa e ar er s res t i m re e ter a aste spa e t ess req e t sta i i i i t r res ts i ess i ter a aste spa e a a sma er e e ti time er ea . O era t ese parameters pr i e a se me a ism a i t e ser (r t e mpi er) t ptimize mem r sa e.

3 5 M B fi

O r i e sta te iq e as a m er a a ta es i terms mem r per rma e. I e era t ese e e ts are a ie e i r i t rea imp eme tati r m er e me a isms t s impr i r a i it t t e i i i a app i ati mem r sa e. mpi er te iq es ma e t is app i ati spe i t i pra ti a .

irst r te iq e ma es prea ati ar e sta s e essar i i t r re es irt a mem r pres s re e r i ar e m ers t rea s. O r a a sis a ie es t is a it t t e se ar pa es i tri te e essar er e r ssi s a irt a mem r aste.

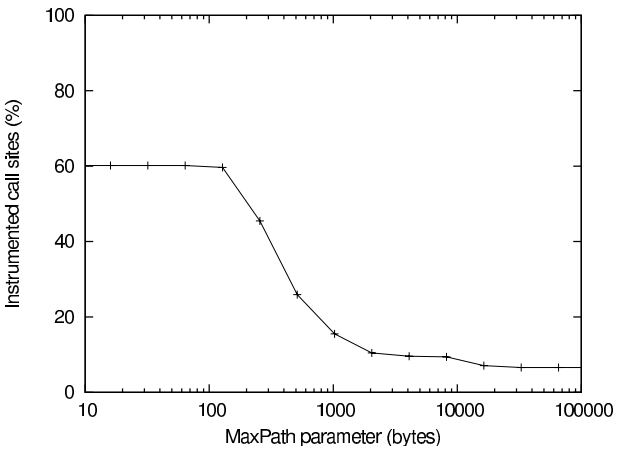
e si i e sta s a impr e pa i e a i r si i a t . Li e sta s are re se i LI O r er i a s sta s t e s are et ee t rea s re i t e size t e app i ati s r i set. A s e a a ate sta s t at are sma er t a a si e pa e t s re i t e era am t mem r aste.

T em strate t e e e t r appr a it respe t t pa i e reate a mi r e mar i i ea t rea repeate a s a ti `bigstack()` i t es a pa es a 1 B er t e sta . T rea s ie et ee a s t `bigstack()`. O r mpi er a a sis i serts a e p i t at t ese a s a t e e p i t a ses a ar e sta t e i e r t e rati t e a . i e `bigstack()` es t ie a t rea s s are a si e 1 B sta it t r sta a a sis e a e t i e ea t rea its i i i a 1 B sta .

e ra t is mi r e mar it 00 t rea s ea i a s `bigstack()` 10 times. e ea t rea as its i i i a sta t e e mar ta es 3.33 se s 1.07 se s i are at ser e e . e si r sta a a sis t e e mar ta es 1.0 se s it 1.00 se s at ser e e s ari a si e sta as rasti a re e t e st pa i . e r i t is test it 1 000 t rea s t e ersi it t r sta a a sis starts t ras i it t e sta a a sis t t e r i time s a es i ear p t 100 000 t rea s.

3 6 S : A h 2 0 4 4

e app ie t is a a sis t t e Apa e 2.0. e ser er. e set t e *MaxPat* parameter t 2 KB t is i e as ma e e ami i t e m er a sites i str me te r ari s parameter a es. T e res ts s i i re 7 i i ate t at 2 KB r KB is a reas a e i e si e ar er parameter a es ma e itt e i ere e i t e era am t i str me tati . e set t e *MinC unk* parameter t KB ase pr i i rmati . B



F u 7: Nu b of A ch 2.0.44 c ll u- d fu c o of h *MaxPath* .

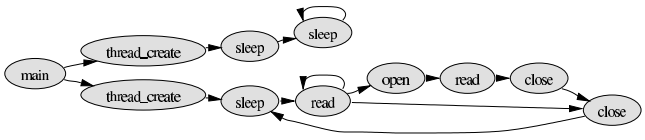
a i pr i ters t e p i ts e etermi e t at i reasi t e size t KB re e t e m er sta i s a i s si i at t rt er i reases ie e a iti a e et. e e pe t t at t ist i met a e a t mate as as t e pr rammer s pp ies a reas a e pr i r a .

Usi t ese parameters est ie t e e a i r Apa e ri e e ti a r a sisti stati e pa es ase t e e 99 e mar s ite. st ti s e e e te e tire it i t e i itia KB e e essar t t rea s i e a e i r er t a a ti t at as a KB er its sta . T is e amp e s s t at e are apa e r i m i e app i ati s it a sma am t sta spa e it t ear sta er .

e ser e t e pr rams e a i rat e a a site r sse ri t e e e ti t is e mar . At 0.1% a sites e p i ts a se a e sta t e i e at a st 27 i str ti s. At 0.5% a sites a ar e sta as i e iti a i r er t a e a e ter a ti sti 20 i str ti s. At 10% a sites a e p i t etermi e t at a e as t req ire i st i str ti s. T e remai i 9% a sites ere a e t e . Ass mi a i str ti s are r eq a i st t e res t is a 73% s e si eri ti a sa e. i e a i str ti s ma e p 5% t e pr rams i str ti s t e era s is appr imate 3% t %.

4 ES CE-AWA E SCHE L

O e t e a a ta es aime r e e t s stems is t at t eir s e i a easi a apt t e app i ati s e e s. e t ase app i ati s are r e i t isti t e e t a ers a mp tati r a parti ar tas pr ee s as t at tas is passe r m a er t a er. T is ar ite t re pr i es t pie es i rmati t at are se r s e i . irst t e rre t a er ra tas pr i es i rmati a t t e tas s ati i t e pr essi ai . T is i rmati a e se t i e pri rit t tas s t at are ser t mp eti e e re i a t es stem. e t e e t s t e a ers tas q e es a e se



F u 8: A x l block h. Th h w d f o u of K o , ou w b v .

t etermi e i sta es are tt e e s a a i i ate e t e ser er is er a e .

apri i pr i es simi ar app i ati spe i s e i r t rea ase app i ati s. i e apri i ses a p erati e t rea i m e e a ie a app i ati as a se q e e sta es ere t e sta es are separate i p i ts. I t is sse se apri i s s e er is q ite simi ar t a e e t ase s stem s s e er. O r met s are m re p er e er i t att e e e t e sta es a t mat i a a e ire t e e t e res r es se ea sta e t s e a i er rai e ami s e i e isi s. I parti ar e se t is a t mate s e i t pr i e a missi tr a t impr e resp se time.

O r appr a a s apri i t pr i es p isti ate app i ati spe i s e i it t req iri t e pr rammer t se mp e r ritt e t i A Is. T s e a impr e per rma e a s a a i it it t mpr misi t e simp i it t e t rea e pr rammi m e.

4 B k h

T e e a stra ti e se r s e i is t e *locking gra* i tai s i rmati a t t e pa es i t e pr ram ere t rea s . a e is a ati i t e pr ram t at e a a e ee ists et ee t es i t e ere se ti e i p i ts. T e ati i t e pr ram is t mere t e a e t e pr ram ter t rat er t e a ai t at as se t rea t e i p i t. T is pat ase appr a a s s t i ere tiate i p i ts i a m re se a t a t e pr ram ter a e a si e t er ise t ere te t e er e s p i ts (e. . t e `read` a `write` s stem a s). i re s s t e i rap r K t a simp e t rea ase e ser er. a t rea a s t is rap i epe e t a e er e t rea is ate at e t ese es.

apri i e erates t is rap at r time ser i t e tra siti s et ee i p i ts. T e e i ea e i t is appr a is t at apri i a *learn* t e e a ir t e app i ati ami a a t e se t at i rmati t impr es e i a a missi tr . T is te iq e r s i part e a se e are tar eti r i pr rams s as I ter et ser ers s it is a epta e t spe time ear i i r er t ma e impr e e isi s ater .

T ma e se t is rap e s e i t rea s e m sta tate t e e es a es it i rmati a t t rea e a i r. T e rst a tati e i tr e is t e a era e r i time rea e e. e at rea s e i e as st tra erse si e e t e pre i s e. e meas re t e time it t t tra erse t e e e si t e e ter a e p ate a e p e tia ei te a era e r t at e e.

e eep a simi ar ei te a era e rea e i e p ate e er time a t rea tra erses e its t i e es. a e s a era e is esse tia a ei te a era e t e e a es si e t e m er p ates is pr p r ti a t t e m er times ea t i e e is ta e. T e e a e t s t e s s t e e t e e i ta e *on average*.

i a e a tate t e a es i res r e sa e. r re t e e e res r es as mem r sta spa e a s ets a e tra t em i i i a . As it U time t ere are ei te a era es r t e es a es. Gi e t at a e t rea is ate at a parti ar e t ese a tati s a s s t estimate et err r i t is t rea i i rease r e rease t e t rea s sa e ea res r e. T is estimate is t e asis r res r e a are s e i : e e t at a res r e iss ar e e pr m te es (a t st rea s) t at re ease t at res r e a em te es t at a q ire t at res r e.

4 2 -A S h

st e isti e e t s stems pri ritize e e t a ers stat i a . A ses i rmati s as e e t a er q e e e t s t ami a t e t e s stem. apri i es e step rt er i tr i t e ti res r e a are s e i . I t is se ti es t se t e e i rap t per rm res r e a are s e i t at is t tra spare t a app i ati spe i .

O r strate rres r e a are s e i as t ree parts:

1. Keep tra res r e ti izati e e s a e i e ami a i ea res r e is at its imit.
2. A tate ea e it t e res r es se its t i e ess e a pre i t t e impa t ea res r e s es e et rea s r m t at e.
3. ami a pri ritize es (a t s t rea s) r s e i ase i rmati r m t e rst t parts.

r ea res r e e i rease ti izati ti it rea es ma im m apa it (s as e t er a a t er res r e) a t e e t r tte a s e i es t at re ease t at res r e. e res r e sa e is e a t t pre ere tia s e e est at s me t at res r e er t e ass mpti t at i s i i rease t r p t. re imp rta t e a res r e is er e e pre ere tia s e e est at re ease t e res r e t a i t tras i .

T is m i ati e se it s me steresis te s t eep t e s stem at t r tte it t t e ris t tras i . A iti a res r e a are s e i pr ie s a at ra r a se siti e rm a missi tr si e tas s ear mp eti te t re ease res r es ereas e tas s a ate t em. T is strate is mp ete a ap ti e i t at t e s e er resp s t a es res r e s mpti e t t t e t pe r ei e a ere a . T e spee a aptati is tr e t e parameters t e e p e tia ei te a era es i r i rap a tati s.

O r imp eme tati res r e a are s e i is q ite strai t r ar . e mai tai separate r q e es r e a e i t e i rap . e peri i a etermi e t e re ati e pri rities ea e ase r pre i ti t eir s seq e t res r e ee s a t e era res r e ti izati t e s stem. O e t e pri rities are e se e t a es stri e s e i a t e e se e t t rea s it i es eq e i r m t e es r q e es. B t t ese perati s are $O(1)$.

A e er i ass mpti r res r e a are s e er is t at res r e sa e is i e t e simi ar r ma tas s at a i p i t. rt ate t is ass mpti seems t i pra ti e. it Apa e re amp e t ere is a m st ariati i res r e ti izati a t e e es t e i rap .

4.2.1 Resources

T e res r es e rre t tra are U mem r a e es ript rs. e tra mem r sa e pr i i r ersi t e `malloc()` ami . e ete t t e res r e imit r mem r at i pa e a t a ti it .

r e es ript rs e tra t e `open()` a `close()` a s. T is te iq e a s s t ete t a i rease i pe e es ript rs i e ie as a res r e. rre t e set t e res r e imit estimati t e m er pe e ti s at i resp se time mps p.

e a as tra irt a mem r sa e a m er t rea s t e t s at prese t. V is tra e t e same a as p si a mem r t t e imit is rea e e e rea s me a s te t res r t ta V a ate (e. . 90% t e a ress spa e).

4.2.2 Pitfalls

e e tere s me i teresti pit a s e imp eme t i apri i s res r e a are s e er. irst etermi i t e ma im m apa it a parti ar res r e a e tri . T e ti izati e e at i t ras i rs te epe s t e r a . r e amp e t e is s s stem a s stai ar m re req ests per se i t e req ests are seq e tia i stea ra m. A iti a res r es a i tera t as e t e V s stem tra es spare is a i t t ree p si a mem r . T em ste e ti es ti e a e ist at r ear si s tras i (s as i pa e a t rates) a t se t ese si st i i ate ma im m apa it .

U rt ate t ras i is t a a sa eas t i t ete t si e it is ara terize a e rease i pr ti e r a a i rease i s stem er ea . ie e a meas re er ea pr ti it is i ere t a app i ati spe i ti . At prese t e attempt t ess att r p t si meas res i e t e m er t rea s reate a estr e a t e m er es pe e a se . A t t is appr a seems s ffi ie t r app i ati s s as Apa e m re mp i ate app i ati s mi t e e t r m a t rea i A I t ata st em t e p i it i rm t e r time s stem a t t eir rre t pr ti it .

App i ati spe i res r es as prese t s me a e es. r e amp e app i ati e e mem r ma a eme t i es res r e a ati a ea ati r m t er time s s tem. A iti a app i ati s ma e e t er i a re s ress as s. O e a ai pr i i a A I t r i t e app i ati a i rm t er time s stem a t its i a res r es ma e a reas a es ti . r simpe ases i e mem r a at rs it ma as e p ssi e t a ie e t is a it t e ep t e mpi er.

4.3 Y P fi

O e pr em t at arises it perati e s e i is t att rea s ma t ie t e pr ess r i a ea t air ess r e e star ati . T ese pr ems are miti ate t s me e te t t e a t t at a t e t rea s are part t e same app i ati a are t ere re m t a tr sti . N et e ess ai re t ie is sti a per rma e pr em t at matters.

Be a se e a tate t e rap ami a it t er i time rea e e it is tri ia t t see e s t at ai e t ie : t eir r i times are t pi a r ers ma it e ar er t a t e a era e e e. O r imp eme tati a s t e s stem perat r t see t e i rap i i e e time req e ies a res r es se se i a USR2 si a t t e r i ser er pr ess.

T ist is er a a e e p rti e a app i ati s t apri i . r e amp e i p rti Apa e e ma p a est at i t ie s ffi ie t te . T is res t is t s rprisi si e Apa e e pe ts t r it preemp ti e t rea s. r e amp e it t r s t t att e `close()` a i ses a s et a s metimes ta e 5ms e e t t e me tati i sists t at it ret r s imme iate e i I/O is se e te . T t is pr em e i sert a iti a ie s i r s stem a i rar e re a ater t e a t a a t `close()`. ie t is s ti es t t e pr em i e era it es a s t rea t e e e i t sma er pie es. A etter s ti (i e a e t et imp eme te) is t se m tip e er e t rea s r r i ser e e t rea s. T is appr a a t e se m tip e pr ess rs a it i e ate ies r m asi a tr a e i perati s s as `close()` a s r pa e a t a i .

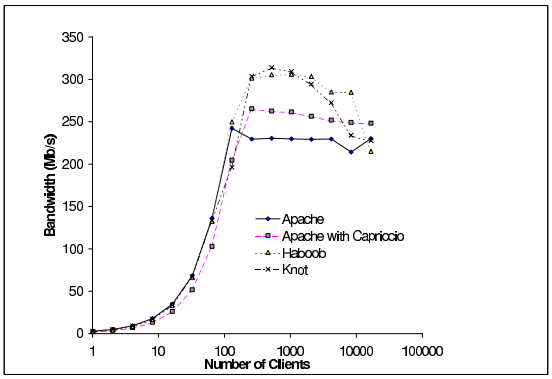
5 EVAL AT

T e mi r e mar s prese te i e ti 2.3 s t at apri i as I/O per rma e a e e e t s a a i i t . I t is se ti e e a ate apri i s per r ma e m re e era er a rea isti e ser er r a . Rea r e r a s i e ar e m ers p te tia s ie ts i pr i e tests t apri i s s a a i i t a s e i . e is ss t e er ea apri i s res r e a are s e er i t is te t a t e is ss t is s e er a e e t a t mati a missi tr .

5 W S v P f

T e ser er ma i e r r e e mar s is a 500 Hz e ti m ser er it 2GB mem r a a I te e1000 Gi a it t er et ar . T e perati s stem is st Li 2. .20. U rt ate e t att e e e pme t series Li er e se i t e mi r e mar s is sse ear ier e ame sta e e p a e er ea a . He e t is e perime t es t ta e a a ta e `epoll` r Li AIO. imi ar e ere ta e t mpare apri i a ai st N TL r t is r a . e ea e t ese a iti a e peri me ts r t re r .

e e erate ie t a it p t 1 simi ar re ma i es a r ss a Gi a it s it e et r . B t apri i a Ha per rm i et r I/O it t esta ar UNIX `poll()` s stem a a se a t rea p r is I/O. Apa e 2.0. (re t se O IX t rea s) ses a m i ati spi p i i i i a e es ript rs a sta ar i I/O a s.



F u 9: W b v b dw d h v u h u b of ul ou cl .

T e r a r t is test siste req ests r 3.2 GB stati e ata it ari s e sizes. T e req est req e ies r ea size a r ea e ere esi e t mat t se t e e 99 e mar . T e ie ts r t is test repeate e t t t e ser er a iss e a series e req ests separate 20ms pa ses.

e imite t e a e sizes Ha a K t t 200 B i r er t r e a ea is a ti it . e se a mi ima rati r Apa e isa i a ami m es a a ess permissi e i . He e it per rme esse tia t e same tas s as Ha a K t.

T e per rma e res ts ere q ite e ra i . Apa es per rma e impr e ear 15% e r er apri i . A iti a K t s per rma e mat e t at t e e e t ase Ha e ser er.

arti ar remar a e is K t s simp i it . K t sists 1290 i es e ritte i a strai t r ar t rea e st e. K t as er eas t rite (it t e s 3 a s t reate) a it is eas t ersta . e si er t is e perie e t e str e i e e r t e sim p i it t e t rea e appr a t riti i rre t app i ati s.

5.2 B k h S

ai tai i i rmati a t t e res r es se at ea i p i t req ires t etermi i ere t e pr ram is e it s a per rmi s me am t mp ta ti t sa e a a re ate res r e ti izati res.

Ta e 2 q a ti est is er ea r Apa e a K t r t e r a es ri e a e. T e t pt i ess t e a era e m er app i ati est at ea app i ati spe t i r m e i p i t t t e e t. T e t t m t i ess t e m er est at apri i spe s i ter a i r er t mai tai i rmati se t e res r e a are s e er. A e ts are t e a era e m er es per i rap e e ri rma pr essi (i.e. er a a a ter t e mem r a e a ra pre i t rs a e arme p).

It is imp rta t t t e t att ese e ts i e *only* t e time spe t i t e app i ati itse . Ker e time spe t

	Item	Cycles	Enabled
Apps	Apache	32697	n/a
	Knot	6868	n/a
System	stack trace	2447	Always for dynamic BG
	edge statistics	673	During sampling periods

T bl 2: Av - d c cl cou fo l c o o C cc o.

I/O pr essi is t i e . i e I ter et app i ati s are I/O i te si e m t eir r a t a ta es pa e i t e er e. He e t e per rma e impa t t is er ea is er t a Ta e 2 s est.

T e er ea at eri a mai tai i statisti s is re ati e sma ess t a 2% r e es i Apa e. re er t ese statisti s te t remai air stea i t e r a s e a e teste s t e a e samp e re ati e i req e t . e a e a samp i rati 1/20 t e q ite s ffi ie t t mai tai a a rate ie t e s stem. T is appr a re es t e a re ate er ea t a mere 0.1%.

T e er ea r m sta tra es is si i a t i er am ti t r % t e e e ti time r Apa e a 3 % r K t. A iti a si e sta tra es are esse tia r etermi i t e ati i t e pr ram t e m st a a s e e a e .

T e er ea r m sta tra i i strates mpi er i te rati e p t impr e apri i s per rma e. T e er ea t mai tai ati i rmati i a stati a e erate i rap is esse tia zer . A t er m re ami te iq e e t mai tai a a aria e t at s a erpri t t e rre t sta . T is erpri t a e p ate at ea ti a XOR i a iq e ti I at ea ti se tr a e it p i t t ese e tra i str ti s a easi e i serte t e m pi er. T is erpri t is t as a rate as a tr e sta tra e t its e a rate e t e erate t e same i rap t at e rre t se.

5 3 -A A C

T test r res r e a are a missi tr a rit ms e reate a simp e s mer pr er app i ati . r er t rea s p a i mem r t a a p a ra m t i pa est r e t em t sta i mem r (r t a se V a ts r pa est at a e ee s appe t). s mer t rea s p rem i mem r r m t e a p a reei it.

T is e mar tests a m er s stem res r es. irst i t e pr ers a ate mem r t q i t e pr ram ma r t irt a a ress spa e. A iti a i pa e t i pr ee st q i t e ma i e i t ras as t e irt a mem r s stem se spa es t a r m is . T e a t e is t ma imize t e tas t r p t (meas re m er pr er ps per se) i e as ma i t e est se t mem r a is res r es.

At r time t e test app i ati is parameterize t e m er s mers a pr ers. R i er Li T rea s i t ere are m re pr ers t a s mers (a te e t ere are e er) t e s stem q i starts t t ras . U er apri i e er t e res r e a are s e er q i ete ts t e er a iti s a im its t e m er pr er t rea s r m r i . T s app i ati s a rea a stea state ear t e ee t e per rma e r e.

6 ELATE W

P o Mod l fo H h Co cu c

T ere as ee a sta i e ate i t e resear mm it a t t e est pr rammi m e r i rre t is e ate as te se t rea s a e e ts i parti ar. O ster t 25 e merate a m er p te tia a a ta es r e e ts. imi ar re e t r s a a e ser er s stems a ates t e se e e ts. am pes i e I ter et ser ers s as as 2 a Har est 10 a ser er i rastr t res i e A 3 a Ni a 35 .

I t e tra iti t e a it ar me t e e pe La er a Nee am 1 e a e pre i s ar e t at a appare t a a ta es e e ts are simp arti a ts p r t rea imp eme tati s 3 . He e e e ie e past ar me ts i a r e e ts are etter ie e as ar me ts r app i ati spe i ptimizati a t e ee r effi ie t t rea r times. B t t ese ar me ts are ma r m ti ati s r apri i . re er t e i rap se apri i s s e er as ire t i spire A sta es a e p i it q e es.

I pre i s r 3 e as prese te a m er reas s t at t rea s s e pre erre er e e ts r i rre t pr rammi . T is paper pr i es a i ti a e i e e r t at aim em strati apri i s per rma e s a a i it a a i it t per rm app i ati spe i ptimizati .

A a e t a . 1 p i te t t at t e e ate et ee e e t ri e a t rea e pr rammi a a t a e sp i t i t e ates: e et ee preempti e a perati e tas ma a eme t a e et ee a t mati a ma a sta ma a eme t. T e i t e term sta rippi t e s ri e t e pr ess ma a sa i a rest ri i e state a r ss i p i ts a t e i e ti t is pr ess as t e primar ra a t ma a sta ma a eme t. T e a t rs as p i t t t e a a ta es t e perati e t rea i appr a .

a a t rs a e attempte t impr e t rea i per rma e tra s rmi t rea e e t e e t ase e. r e amp e A a e t a . 1 a t mate t e pr ess sta rippi i e e t ri e s stems a i e t e ritte i a m re t rea i e st e. I s me se se t a t rea pa a es per rm t is same tra s ati at r time mappi i perati s i t i state ma i es er eat . U timate e e ie e t ere is a a ta e t a stati tra s rmati r m t rea e e t e e t ri e e e a se a e t e t rea r time a per rm st as e as a e e t ase e. O r per r ma e tests it apri i rr rate t is aim.

U -L v l Th d

T ere a e ee ma ser e e t rea pa a es t t e i er r m apri i i t eir as a te iq es. T t e est r e e apri i is iq e i its se t e i rap t pr i e res r e a are s e i a i its se mpi e time a a sis t e e t app i ati spe i ptimizati s. A iti a e are t a are a a a e i epe e t t rea i i rar t at ses i e sta rames t e is ss s me a a e epe e t es e .

i ame ts 2 a NT s i ers are t i per rma e ser e e t rea pa a es. B t se perati e s e i t t e are t tar ete at ar e m ers i t rea s. i ima te t it i T rea s 1 is a

i per rma e t rea pa a e spe ia ize r e a es t at i es ast is i raries a mem r ma a eme t. T e per rma e ptimizati s emp e t ese pa a es e se r apri i as e t ese are mp eme tar t r r .

T e tate T rea s pa a e 3 is a i t ei t pera ti e t rea i s stem t at s ares apri i s a simp i i t e pr rammi m e r et r ser ers. U i e apri i t e tate T rea s i rar es t pr i e a O IX t rea i i ter a e s app i ati s m st e re rit te t se it. A iti a tate T rea s se eit er `select` r `poll` i stea t e m res a a e Li `epoll` a t e se i is I/O. T ese a t rs imit t e s a a i it tate T rea s r et r i te si e r a s a t e restri t its rre r is i te si e r a s. T ere are pat es a ai a e t a Apa e t se tate T rea s 33 res ti i a per rma e i rease. T ese pat es i e a m er t er impr eme ts t Apa e e er s it is imp ssi e t te m t e im pr eme t ame r m tate T rea s. U rt ate t ese pat es are er mai tai e a t mpi e ea s e ere a e t r ire t mparis s a ai st apri i .

e era ti ati s 2 s et e pr em i I/O a e pe te i /preempti ser e e t rea s a i er e s pp rt r ti i t e ser e e s e er t ese e e ts. T is appr a e s res ea i te rati t e t rea i rar a t e perati s stem e er t e ar e am t er e a es i e seem t a e pre e i e a pti . A t er p te tia pr em it t is appr a is t at t ere i e e s e era ti ati r ea tsta i I/O perati i a m er i t e te s t sa s r I ter et ser ers. T is res t is trar t t e ri i a a re i t e m er er e t rea s ee e . T is pr em appare t stems r m t e a t t at s e era ti ati s are e e pe primari r i per rma e mp ti e ir me ts ere is a ast et r I/O are mi a t. Ne ert e ess s e er a ti ati s a e a ia e appr a t ea i it pa e a ts a preempti s i apri i . mp i s e er a ti ati s as a t e ser e e s e er t i e e t e er e s e isi a t i er e t rea t preempt. T is s eme a e se t s e iffi t pr ems i e *riority inversion* a t e *convoy eno enon* .

pp rt r ser e e preempti a :N t rea i (i.e. r i ser e e t rea s t p N er e t rea s) is tri . Te iq es s as ptimisti rre tr a i s r steai 7 a e se e e ti e t ma a e t rea a s e er ata str t res. r i a prese ts a i e es ripti t ese a t er te iq es i t e te t Li 12. e e pe t t emp ma t ese te iq es i apri i e e a s pp rt r :N t rea i .

K l Th d

T e N TL pr e t r Li as ma e reat stri es t ar impr i t e effi ie Li er e t rea s. T ese a a es i e a m er er e e e impr eme ts s as etter ata str t res er mem r er ea a t e se $O(1)$ t rea ma a eme t perati s. N TL is q ite e a is sti era ti e e e pme t. He e e e pe t t at s me t e per rma e e ra ati e it i er m ers t rea s ma e res e as t e e e pers s a reate aster a rit ms.

A l c o - c fic z o

er rma e ptimizati t r app i ati spe i tr s stem res r es is a imp rta t t eme i O re sear . a 21 a e app i ati s t spe i t eir V pa i s eme i impr e per rma e r ap p i ati s t at e a t t eir p mi mem r ee s a is a ess patter s. UN T 37 i simi ar t i s r et r I/O impr i e i i it a re i er ea it t mpr misi sa et . T e IN perati s stem 5 a t e VINO perati s stem 29 pr i e ser st mizati a i app i ati et em e i t t e er e . T e er e 13 t t e pp site appr a a m e m st t e O t ser e e . A t ese s stems a app i ati spe i ptimizati ear a aspe ts t e s stem.

T ese te iq es req ire pr rammers t tai r t eir ap p i ati t ma a e res r es r itse t is t pe t i is te iffi t a ritt e. A iti a t e tie pr rams t sta ar A Is re i t eir p rta i it . apri i ta es a e appr a t app i ati spe i ptimizati e a i a t mati mpi er ire te a ee a ase t i t e t rea pa a e. e e ie e t att is appr a i ma e t ese te iq es m re pra ti a a i a a i er ra e app i ati s t e e t r m t em.

A ch o ou I/

A m er a t rs pr p se impr e er e i ter a es t at a e a imp rta t impa t ser e e t rea i . As r s I/O primiti es s as Li s `epoll` 20 is AIO 17 a reeB s q e e i ter a e 19 are e tra t reati as a a e ser e e t rea pa a e. apri i ta es a a ta e t ese i ter a es a e e t r m impr eme ts s as re i t e m er er e r ssi s.

ck M

T ere are a m er re ate appr a es t t e pr em prea ati ar e sta s. me ti a a a es s as ta ar L Ne Jerse 3 t se a a sta at a rat er t e a ate a a ti ati re r s t e eap. T is appr a is reas a e i t e te t a a a e t at ses a ar a e e t r a t at s pp rts i er r er ti s a rst ass ti ati s . H e er t ese eat res are t pr i e t e pr rammi a a e i mea s t at ma t e ar me ts i a r eap a ate a ti ati re r s t app i r ase. rt erm re e t is t i r t e er ea ass i ate it a i a ar a e e t r t r s stem pre i s r ass t at Ja a s e era p rp se ar a e e t r is i appr priate r i per rma e s stems 30.

A m er t er s stems a e se ists sma sta si p a e ti s sta s. B r a e reit es ri e a te iq e t at ses a si e sta r m tip e e ir me ts e e ti e i i i t e sta i t s sta s e er t e t a a ze t e pr ram t attempt t re e t e am t r time e s req ire . O e i is a a a e a r time s stem r para e izi pr rams se a simp i e ersi B r a e reit s te iq e a e spa etti sta s 9. I t is te iq e a ti ati re r s r i ere t t rea s are i ter ea e a si e sta e er ea a ti ati re r s i t e mi e t e sta a t e re aime i i e a ti ati re r s sti e ist rt er t e sta i a a t e am t aste sta spa e t r it t .

re re e t t e Laz T rea s pr e t i tr e sta ets i are i e sta s r se i mpi i para e a a es 15 . T is me a ism pr i es r time sta er e s a it ses a mpi er a a sis t e imi ate e s e sta sa e a e e e er t is a a sis t at es t a e re rsi as apri i es a it es t pr i e t i parameters. e a B e as se e size sta ets t pr i e s pr essi time i a para e rea time ar a e e t r 11 .

7 F T E W

e are i t e pr ess e te i apri i t r it m ti U ma i es. T e ame ta a e e pr i e m tip e Us is t at e a er re t e p erati e t rea i m e t pr i e at mi it . H e er e e ie e t at i rmati pr e t e mpi er a assist t e s e er i ma i e isi s t at ara tee at mi it ertai s e att e app i ati e e.

T ere are a m er aspe ts apri i s imp eme ta ti e i e t e pr re. e e ie e e ramati a re e er e r ssi s er ea et r a it a at i i ter a e ras r s et r I/O. e as e pe t t ere are ma a s t impr e r res r e a are s e er s astra i t e aria ei t e res r e sa e i rap es a impr i r ete ti t ras i .

T ere are se era a s i i r sta a a sis a e impr e . As me ti e ear ier e se a ser ati e appr imati t e a rap i t e prese e ti p i ters r t er a a e eat res t at req ire i ire t a s (e. . i er r er ti s irt a met ispat a e epti s). Impr eme ts t t is appr imati s sta tia impr e r res ts. I parti ar e pa t a apt t e ata a a sis re 2 i r er t isam i ate ma t e ti p i ter a sites. e mpi i t er a a es e start it simi ar ser ati e a rap s a t e emp e isti tr a a ses (e. . t e 0 A a a ses 31 r ti a a e t rie te a a es a a es r irt a ti res ti a a ses 27 r e t rie te a a es).

I a iti e p a t pr e pr i t s t at a assist t e pr rammer a t e mpi er i t i apri i s sta parameters t e app i ati s ee s. I parti ar e a re r i rmati a t i ter a a e ter a aste spa e a e a at er statisti s a t i ti a s a se e sta s t e i e . B ser i t is i rmati r a ra e parameter a es e a a t mate parameter t i . e a a s s est p te tia ptimizati s t t e pr rammer i i ati i ti s are m st te resp si e r i reasi sta size a sta aste.

I e era e e ie e t at mpi er te i p a a imp rta t r e i t e e ti t e te iq es e s ri e i t is paper. r e amp e e are i t e pr ess e isi a mpi er a a sis t at is apa e e erati a i rap at mpi e time t ese res ts i impr e t e effi ie t er time s stem (si e a tra es are req ire t e erate t e rap) a t e i a s t et at mi it r ree ara tee i stati a t at ertai riti a se ti s t tai i p i ts. I a iti e p a t i esti ate strate ies r i serti i p i ts i t t e e at mpi e time i r er t e r e air ess.

mpi e time a a sis a a s re e t e rre e s ar i t e pr rammer a t ata ra es. A t stati ete ti ra e iti s is a e i t ere as ee re e t pr ress e t mpi er impr eme ts a tra ta e e pr ram a a ses. I es 1 a a a e r et r e se s rs t ere is s pp rt r at mi se ti s a t e mpi er ersta s t e rre m e. It ses a mi t re I/O mp eti s a r t mp eti t rea s a t e mpi er ses a ariati a a rap t at is simi ar t r i rap . T e mpi er e s res t at at mi se ti s resi e it i e e e t at rap i parti ar a s it i a at mi se ti a t ie r (e e i ire t). T is i s pp rt e e treme p er r a t ri ser ers. i a e e pe t t at at mi se ti s i as e a e etter s e i a e e ea ete ti .

8 C L S S

T e apri i t rea pa a e pr ies empiri a e i e e t at i t rea pa a es is a ia e s ti t t e pr em i i s aa e i rre I ter et ser ers. O re perie e it riti s pr rams s ests t att e t rea e pr rammi m e is a m re se a stra ti t a t e e e t ase m e r riti mai tai i a e i t ese ser ers. B e pi t e t rea imp e me tati r m t e perati s stem itse e a ta e a a ta e e I/O me a isms a mpi er s pp rt. As a res t e a se te iq es s as i e sta s a res r e a are s e i i a s t a ie e si i a t s aa i it a per rma e impr eme ts e mpare t e isti t rea ase r e e t ase s stems.

As t is te mat res e e pe te e m re t ese te iq es t e i te rate it mpi er te . B riti pr rams i t rea e st e pr rammers pr i e t e mpi er it m re i rmati a t t e i e e str t re t e tas s t at t e ser er m st per rm. Usi t is i rmati e pe t at t e mpi er a e p se e e m re pp rtities r t stati a ami per rma e t i .

9 EFE E CES

1 A. A a J. H e . T eimer . J. B s a J. R. e r. perati e tas ma a eme t it t ma a sta ma a eme t. I *Proceedings of t e 2002 Usenix ATC* J e 2002.

2 T. . A ers B. N. Bers a . . Laz s a a H. . Le . e er A ti ati s: e ti e Ker e pp rt r t e User Le e a a eme t ara e ism. *ACM Transactions on Co uter Syste s* 10(1):53–79 e r ar 1992.

3 A. . Appe a . B. a Q ee . ta ar L Ne Jerse . I *Proceedings of t e 3rd International Sy osiu on Progra ing Language I le entation and Logic Progra ing* pa es 1–13 1991.

A. . Appe a Z. a . A empiri a a a ti st sta s. eap st r a a es it s res. *Journal of Functional Progra ing* (1): 7–7 Ja 199 .

5 B. N. Bers a . am ers . J. ers . ae a . Namee . ar a . a a e a . G. irer. IN a e te si e mi r er e r

app i ati spe i perati s stem ser i es. I *ACM SIGOPS Euro ean Works o* pa es –71 199 .

. .B as e J. Gra . . it ma a T. G. ri e. T e p e me . *O erating Syste s Revie* 13(2):20–25 1979.

7 R. .B m e . .J er B. .K szma . . Leisers K. H. Ra a a Y. Z . i : A effi ie t m tit rea e r time s stem. *Journal of Parallel and Distri uted Co uting* 37(1):55– 9 199 .

. G. B r a B. e reit. A m e a sta imp eme tati m tip e e ir me ts. *Co unications of t e ACM* 1 (10):591– 03 O t 1973.

9 . . ar is e A. R ers J. Repp a L. He re . ar e perie es it O e . I *Proceedings of t e 6t International Works o on Languages and Co ilers for Parallel Co uting (LNCS)* 1993.

10 A. a t . B. a zi . Neer ae s . . artz a K. J. rre . A Hierar i a I ter et O e t a e. I *Proceedings of t e 1996 Usenix Annual Tec nical Conference* Ja ar 199 .

11 . e a G. .B e . A para e rea time ar a e e t r. I *Proceedings of t e 2001 ACM SIGPLAN Conference on Progra ing Language Design and I le entation (PLDI '01)* 2001.

12 J. r i a. ast m tit rea i s are mem r m tipr ess rs. Te i a rep rt U i ersit a ta J e 2000.

13 . R. er . .Kaas e a J. O T e. er e: A perati s stem ar ite t re r app i ati e e res r e ma a eme t. I *Sy osiu on O erating Syste s Princi les* pa es 251–2 1995.

1 . Ga . Le is R. Be re . e s . Bre er a . er. T e es a a e: A isti appr a t et r e em e e s stems. I *ACM SIGPLAN Conference on Progra ing Language Design and I le entation* 2003.

15 . .G stei K. . a ser a . . er. Laz T rea s ta ets a r izers: a i primiti es r mpi i para e a a es. I *T ird Works o on Langauges, Co ilers, and Run-Ti e Syste s for Scala le Co uters* 1995.

1 T. H . i ima te t T rea 0.7 ma a . ttp:// .ara et r . m/ s/m t ma a .p 2002.

17 B. LaHaise. Li AIO me pa e. ttp:// . a . r / a /ai /.

1 H. .La er a R. .Nee am. O t e a it perati s stem str t res. I *Second Inernational Sy osiu on O erating Syste s, IR1A* O t er 197 .

19 J. Lem . Kq e e: A e eri a s a a e e e t ti ati a i it . I *USENIX Tec nical conference* 2001.

20 . Li e zi. Li ep pat . ttp:// . mai ser er. r / i pat es/ i impr e. tm .

21 . Namee a K. Armstr . te i t e a e ter a pa er i ter a e t a mm ate ser e e pa e rep a eme t p i ies. Te i a Rep rt TR 90 09 05 U i ersit as i t 1990.

22 . . i . *Advanced Co iler Design and I le entation.* r a Ka ma a ra is 2000.

23 G. .Ne a . ea . .Ra a . eimer. IL: I terme iate a a e a t s r a a sis a tra s rmati pr rams. *Lecture Notes in Co uter Science* 230 :213–229 2002.

2 G. .Ne a . ea a . eimer. re : T pe sa e retr tti e a e. I *T e 29t Annual ACM Sy osiu on Princi les of Progra ing Languages* pa es 12 –139. A Ja . 2002.

25 J. K. O ster t. T rea s Are A Ba I ea (r m st p rp ses). rese tati i e at t e 199 Use i A a Te i a ere e Ja ar 199 .

2 V. . ai . r s e a . Z ae ep e . as : A ffi ie t a rta e e er er. I *Proceedings of t e 1999 Annual Usenix Tec nical Conference* J e 1999.

27 H. . a e a B. G. R er. ata ase irt a ti res ti . *Lecture Notes in Co uter Science* 11 5:23 –25 199 .

2 . a a . .G i . A a rit m r s i strai t satis a ti pr ems.

29 . I. e tzer Y. . ma a K. A. mit . ea i it isaster: r i i mis e a e er e e te si s. I *Proceedings of t e 2nd Sy osiu on O erating Syste s Design and I le entation* pa es 213–227 eatt e as i t 199 .

30 . A. a . a e . J. ra i a J. . He erstei . Ja a s pp rt r ata i te si e s stems: perie es i i t e Te e rap ata s stem. *SIGMOD Record* 30( ):103–11 2001.

31 O. i ers. *Control-Flo Analysis of Hig er-Order Languages.* t esis ar e ie e U i ersit a 1991.

32 . T er i . r ti e i rar s r e. ttp:// . r . e/~ r ese/ r /.

33 U . A e erati Apa e pr e t. ttp://aap.s r e r e. et/.

3 U . tate t rea s r I ter et app i ati s. ttp://state t rea s.s r e r e. et/ s/st. tm .

35 J. R. Be re . Bre er N. B ris . e . e s J. a a J. La . Gri e a . er. Ni a: A rame r r et r ser i es. I *Proceedings of t e 2002 Usenix Annual Tec nical Conference* J e 2002.

3 R. Be re J. it a . Bre er. e e ts are a a i ea (r i rre ser ers). I *Proceedings of t e 2003 HotOS Works o* a 2003.

37 T. i e A. Bas V. B a . V e s. U Net: A User Le e Net r I ter a e r ara e a istri te mp ti . I *Proceedings of t e 15t ACM Sy osiu on O erating Syste s Princi les* pper tai Res rt O U A e eme er 1995.

3 . e s . . er a . A. Bre er. A: A ar ite t re r e iti e s a a e I ter et ser i es. I *Sy osiu on O erating Syste s Princi les* pa es 230–2 3 2001.