Application BASTRI

Fiches Equipes

PACAP (SR0750NR)

Pushing Architecture and Compilation for Application Performance
ALF (SR0460XR) PACAP

Statut: Décision signée

Responsable : Erven Rohou

Mots-clés de "A - Thèmes de recherche en Sciences du numérique - 2023" : A1.1.1. Multi-cœurs, pluri-coeurs , A1.1.2. Accélérateurs matériels (GPGPU, FPGA, DSP, etc.) , A1.1.3. Modèles memoire , A1.1.8. Sécurité des architectures , A1.1.11. Architectures quantiques , A1.6. Efficacité énergétique , A2.2.1. Analyse statique , A2.2.3. Gestion mémoire , A2.2.4. Architectures parallèles , A2.2.5. Environnements d'exécution , A2.2.6. GPGPU, FPGA... , A2.2.7. Compilation adaptative , A2.2.8. Génération de code , A2.2.9. Sécurité par la compilation , A2.3. Systèmes embarqués et cyber-physiques , A2.3.1. Systèmes embarqués , A2.3.2. Systèmes cyber-physiques , A2.3.3. Systèmes temps réel , A4.4. Sécurité des équipements et des logiciels , A5.10.3. Planification , A5.10.5. Interactions (avec l'environnement, des humains, d'autres robots , A9.2. Apprentissage

Mots-clés de "B - Autres sciences et domaines d'application - 2023" : B1. Sciences du vivant , B2. Santé , B3. Environnement et planète , B4. Energie , B5. Industries du futur , B5.7. Fabrication 3D , B6. Informatique et télécommunications , B7. Transport et logistique , B8. Villes et territoires intelligents , B9. Société & connaissance

Domaine : Algorithmique, programmation, logiciels et architectures
Thème : Architecture, langages et compilation

Période : 01/07/2016 -> 31/12/2025
Dates d'évaluation : 19/03/2020

Etablissement(s) de rattachement : U. RENNES
Laboratoire(s) partenaire(s) : IRISA (UMR6074)

CRI : Centre Inria de l'Université de Rennes
Localisation : Centre Inria de l'Université de Rennes
Code structure Inria : 031081-2

Numéro RNSR : 201622151M
N° de structure Inria: SR0750NR

Présentation

PACAP: Performance des Applications par la Compilation et l'Architecture des Processeurs

En bref, l'objectif à long terme de l'équipe projet PACAP s'articule autour de la *performance* des ordinateurs, c'est-à-dire la vitesse à laquelle les programmes s'exécutent. Nous avons l'intention de contribuer aux besoins croissants de performance, et de garanties de temps d'exécution.

Traditionnellement, le terme « performance » est compris comme « combien de temps il faut pour exécuter un programme ». Il s'agit alors de minimiser la *latence*, le temps moyen d'exécution. Nous nous intéressons aussi à d'autres définitions de la performance. Le *débit* mesure combien de calculs peuvent être effectués par unité de temps. C'est une mesure plus pertinente pour les processeurs massivement parallèles et les GPU qui disposent d'un très grand nombre de cœurs, et pour lesquels la latence est moins critique. Finalement, nous nous intéressons aussi au *pire temps d'exécution* (WCET, pour *worst-case execution time*). Il est très important pour les systèmes temps réel critiques pour lesquels les concepteurs doivent garantir que les bornes de temps d'exécution sont strictement respectées, en toute situation.

La complexité des systèmes actuels est telle que simplement s'assurer de leur bonne performance est une tâche difficile, à laquelle nous allons nous attaquer.

Nous considérons occasionnellement d'autres métriques reliées à la performance, comme l'efficacité énergétique, l'énergie consommée, la complexité globale du système. Notre but est de proposer des solutions qui conduisent à des systèmes informatiques plus efficaces, en prenant en compte les applications, les compilateurs, les systèmes d'exploitation et les microarchitectures — actuels et à venir. Et comme il est d'usage qu'un gain en

Contact

- **Responsable :** Erven Rohou
- **Tél :**
- **Secrétariat Tél :** +3.32.99.84.72.45

En savoir plus

- Site de l'équipe
- Site sur inria.fr
- Site du responsable
- Derniers Rapports d'Activité : 2016 , 2017 , 2018 , 2019 , 2020 , 2021 , 2022 , 2023

Documents sur la structure

- Intranet
- Privés

Décisions

- 11707 (29/06/2016) : création
- 14582 (09/12/2020) : prolongation
- 15177 (13/12/2021) : prolongation

Localisation

- **Adresse postale :** Centre Inria de l'Université de Rennes 263, avenue du Général Leclerc Campus universitaire de Beaulieu 35042 Rennes Cedex France
- **Coordonnées GPS :** 48.116, -1.64

performance s'accompagne d'une dégradation d'un autre aspect, PACAP s'intéresse aux divers compromis.

La décennie passée a vu la fin de la croissance exponentielle de la fréquence d'horloge, et l'introduction des processeurs multicœurs grand public. PACAP connaît la fin de la *loi de Moore* (qui énonce que le nombre de transistors qui composent un circuit double approximativement tous les deux ans) et la généralisation des processeurs massivement parallèles hétérogènes. Cela aura un impact sur la façon dont la performance continue d'augmenter, et dont elle peut être garantie. C'est aussi le moment où de nouveaux paramètres doivent être pris en compte :

1. l'existence de fautes, dont l'impact devient de plus en plus important, à mesure que la taille de gravure diminue ;
2. les besoins en sécurité, à tous les niveaux des systèmes informatiques ;
3. le *green computing*, c'est-à-dire l'amélioration de l'efficacité énergétique.

Axes de recherche

Nous nous efforçons d'améliorer la performance d'une manière aussi transparente que possible pour les utilisateurs. Par exemple, plutôt que de proposer un nouveau langage, nous considérons les applications existantes (écrites par exemple en C), et nous concevons des optimisations de compilation qui profitent immédiatement aux programmeurs; de même, nous proposons de nouveaux mécanismes microarchitecturaux par opposition à des évolutions des jeux d'instructions; nous analysons et ré-optimisons des programmes au format binaire automatiquement, sans aucune intervention de l'utilisateur ou du programmeur.

Le périmètre des recherches de l'équipe-projet PACAP découle de l'intersection de deux axes: d'une part nos objectifs à long-terme, conditionnés par le contexte général des systèmes informatiques ; d'autre part l'expertise et le savoir-faire des membres de l'équipe.

Performance orientée « latence »

Améliorer le temps d'exécution moyen des systèmes informatiques a été le cœur de métier des équipes CAPS et ALF depuis deux décennies. Nous prévoyons de poursuivre cette direction de recherche, en intervenant à tous les niveaux : compilation, optimisations dynamiques et microarchitecture.

Performance orientée « débit »

Le but est de maximiser le ratio performance-puissance consommée. Nous allons tirer parti du modèle d'exécution des architectures orientées débit (comme les GPU) pour l'étendre vers des systèmes à usage général. Pour nous attaquer au problème des accès mémoire (le *memory wall*), nous envisageons des techniques d'économie de bande passante, telles que les hiérarchies mémoire et la compression de la mémoire.

Systèmes temps réel — WCET

Les concepteurs de systèmes temps-réel doivent fournir une limite supérieure au pire temps d'exécution (WCET : *worst-case execution time*) des tâches de leurs systèmes. Par définition, cette borne doit être sure (c'est-à-dire supérieure à tout temps d'exécution possible). Pour être utiles, les estimations de WCET doivent aussi être aussi précises que possible. Le processus d'obtention d'une borne de WCET consiste à analyser un exécutable au format binaire, à modéliser le matériel, puis à maximiser une fonction objective qui prend en compte tous les chemins d'exécution possibles et leurs temps d'exécution respectifs. Nos recherches considèrent les directions suivantes: 1) mieux modéliser le matériel pour soit améliorer la précision, soit traiter des architectures plus complexe (par exemple multicœur); 2) éliminer les chemins infaisables de l'analyse; 3) envisager des approches probabilistes où les estimations de WCET sont fournies avec un niveau de confiance.

Évaluation de la performance

La loi de Moore décrit l'évolution de la complexité de la micro-architecture des processeurs, qui se répercute sur toutes les autres couches : hyperviseurs, systèmes d'exploitation, compilateurs et applications suivent des tendances similaires. Bien qu'une petite catégorie d'experts soit en mesure de comprendre (en partie) le comportement d'un système, la grande majorité des utilisateurs ne sont exposés qu'à – et intéressés par – le résultat net : à quelle vitesse leurs applications s'exécutent réellement. En présence de machines virtuelles et de *cloud*, les charges de travail multiples viennent ajouter encore au non-déterminisme de la performance. Nous prévoyons d'étudier comment les performances des applications peuvent être caractérisées et présentées à un utilisateur final : comportement de la microarchitecture, métriques pertinentes, éventuellement rendu visuel. Ciblant notre propre communauté, nous allons étudier des moyens rapides et précis pour simuler les architectures du futur, y compris hétérogènes.

Une fois diagnostiqués, la façon dont les goulots d'étranglement sont traités dépend du niveau d'expertise des utilisateurs. Les plus experts peuvent généralement utiliser directement un diagnostic et sont à même de résoudre leur problème. Les utilisateurs moins qualifiés doivent être guidés vers la meilleure solution. Nous prévoyons de nous appuyer sur la compilation itérative pour générer plusieurs versions des régions de code critiques, utilisables

chacune dans certaines conditions d'exécution. Pour éviter l'explosion de la taille de code résultant du *multiversioning*, nous allons tirer parti de la *split-compilation* pour intégrer des « recettes » de génération de code à appliquer juste-à-temps, ou même pendant l'exécution grâce à la traduction binaire dynamique. Enfin, nous allons explorer l'applicabilité de l'*autotuning*, où les programmeurs exposent des paramètres de leur code qui peuvent être modifiés pour générer des versions alternatives du programme (comme un compromis entre l'énergie consommée et la qualité de service) et laissent un orchestrateur global prendre les décisions.

Traitement des fautes — Fiabilité

L'évolution de la technologie des semi-conducteurs suggère que le taux de fautes permanentes va augmenter de façon spectaculaire avec la diminution de la taille de gravure. Alors que de nombreuses techniques existent pour corriger les fautes (comme les codes correcteurs d'erreurs) et fournir des circuits sans défaut, la croissance exponentielle du nombre de fautes les rendra inabordables dans l'avenir. Par conséquent, d'autres approches, comme la désactivation à grain fin et la reconfiguration d'éléments matériels (par exemple des unités fonctionnelles individuelles ou des blocs de cache) deviendra économiquement nécessaire. Cette désactivation fine dégradera les performances par rapport à une exécution sans faute. Cette évolution a un impact sur la performance (à la fois moyenne et pire cas). Nous prévoyons d'aborder cette évolution, et de proposer de nouvelles techniques, qui peuvent être développées à tous les niveaux. Par exemple, au niveau de la microarchitecture, on pourrait envisager de concevoir une partie d'un cache dans une ancienne technologie pour garantir un niveau minimal de performance ; à la compilation, on pourrait générer un code redondant pour les sections critiques ; au moment de l'exécution, on peut détecter les fautes et appliquer des mesures correctives au logiciel ou au matériel. Des solutions mixant plusieurs niveaux sont également très prometteuses.

Traitement des attaques — Sécurité

Les systèmes informatiques sont en permanence soumis aux attaques, de jeunes pirates qui tentent de montrer leurs compétences, à des criminels « professionnels » qui tentent de voler des informations de carte de crédit, et même des agences gouvernementales disposant de ressources pratiquement illimitées. Une grande quantité de techniques ont été proposées dans la littérature pour contourner ces attaques. Beaucoup d'entre elles causent des ralentissements importants en raison des contrôles supplémentaires et des contre-mesures additionnelles. Grâce à notre expertise des techniques de microarchitecture et de compilation, nous sommes en mesure d'améliorer considérablement l'efficacité, la robustesse et la couverture des mécanismes de sécurité, ainsi que de travailler en partenariat avec des experts pour concevoir des solutions innovantes.

Efficacité énergétique

La consommation électrique est devenue une préoccupation majeure des systèmes informatiques, quelque soit le facteur de forme, allant de capteurs autonomes pour l'Internet des objets, à des systèmes portables embarqués alimentés par batterie, et jusqu'à des superordinateurs consommant des dizaines de mégawatts. Temps d'exécution et énergie sont souvent des objectifs corrélés. L'optimisation de performance sous contrainte de puissance, cependant, introduit de nouveaux défis. Il apparaît aussi que les technologues introduisent de nouvelles solutions (par exemple la RAM magnétique) qui, à son tour, entraînent de nouveaux compromis et des opportunités d'optimisation.

Relations industrielles et internationales

PACAP est impliquée dans les collaborations suivantes. Pour de plus amples détails, veuillez consulter la page publications du site web.

- CONTINUUM: Design Continuum for Next Generation Energy-Efficient Compute Nodes – projet ANR, oct 2015 à avr 2019.
- SECODE: Secure Codes to thwart Cyber-physical Attacks, ANR CHIST-ERA, jan 2016 à déc 2018.
- ANTAREX: AutoTuning and Adaptivity appRoach for Energy efficient eXascale HPC systems, H2020 FET HPC, sep 2015 à aoû 2018.
- W-SEPT: “WCET: SEmantics, Precision and Traceability” – projet ANR.
- Capacites: Calcul Parallèle pour Applications Critiques en Temps et Sûreté – Parallel computations for safty critical real-time applications, PIA projet “Investissement d’avenir”, oct 2014 à fév 2018.
- Large scale multicore virtualization for performance scaling and portability – Inria Project Lab.
- Nano2017 – PSAIC (Performance and Size Auto-tuning through Iterative Compilation) – Programme de recherche & développement coopératif – Inria / STMicroelectronics
- ARGO (WCET-Aware Parallelization of Model-Based Applications for Heterogeneous Parallel Systems) projet H2020.
- ZEP (ZEro Power computing systems): Inria Project Lab, 2017 à 2020.
- DYVE (Dynamic vectorization for heterogeneous multi-core processors with single instruction set): ANR JCJC, avr 2020 à oct 2023.
- NOPE (Normally-Off Platforms for Embedded Systems): action exploratoire du LabEx Cominlabs, jan à déc 2019.
- ARMOUR (dynAmic binaRy optiMizatiOn cyber-secURity): 2018 à 2021.
- Hybrid SIMD architectures: 2018 à 2019.

- LabEx CominLabs *NOP: Safe and Efficient Intermittent Computing for a Batteryless IoT*, oct 2021 à déc 2024.
- Projet ANR (ANR-21-CE25-0016-02) *Maplurinum: Machinæ pluribus unum*, oct 2021 à sep 2025.
- Action exploratoire Inria Ofast3D: Compilateur optimisant pour impression 3D rapide, oct 2021 à sep 2025.

Nous sommes membres de :

- Réseau d'excellence HiPEAC: European Network of Excellence on High Performance and Embedded Architecture and Compilation
- COST action TACLe: Timing Analysis at Code Level