What Good is a Scheduling Competition? - Insights from the IPC

Terry Zimmerman

Carnegie Mellon University, Robotics Institute
5000 Forbes Avenue, Pittsburgh, PA
wizim@cs.cmu.edu

Abstract

The intention of this paper is twofold: 1) Review the International Planning Competition events that have been held to date with respect to their design, features that worked well and less well, and their reported evaluation results, and 2) Direct attention to those aspects and insights based on IPC experience that are most relevant to the possible design of a scheduling competition, and indeed whether the evidence supports development of an IPC-like competition for the scheduling community.

Introduction

The impetus to design and conduct a scheduling competition at this time is likely gaining support in part due to the accumulated experience with the International Planning Competition (IPC). The IPC, first hosted in 1998 at AIPS-98, provides a useful historical and on-going example of research community response to a structured competition for developers of state-of-the-art A.I. systems. The competition has been held semi-annually in conjunction with AIPS / ICAPS conferences, with the most recent being IPC-5 at the 2006 ICAPS conference.

There are important differences in the 'maturity' level of planning and scheduling in terms of fielded applications but the nature of the two technologies, their close interrelationships, and the active participation levels of the two research communities suggests that the IPC experience may be a reasonable predictor of the desirability and the direction a scheduling competition should take. We examine here the track record of the five IPC events held to date with the goal of assessing what features, successes and failures might suggest about the design and value of a comparable scheduling competition.

The paper is organized as follows: The next section presents the stated goals of the IPC and contains a subsection assessing each in turn. The discussion here focuses on both the effectiveness of the events in actually achieving the stated goal and the relevance of the goal to a potential scheduling competition. Were appropriate we suggest features that we deem particularly significant for exploring and advancing the state of the art in scheduling. This is followed by a section summarizing the relevance of the IPC experience to a possible scheduling competition and our recommendations. Finally, an appendix is provided featuring a synopsis of each of the five IPC events that have been staged to date from which this paper's observations and assessments have been drawn.

IPC Goals and Their Relevance to a Scheduling Competition

The stated general goals of the IPC are :[1]

1. analyzing and advancing the planning state-of-the-art
2. providing new benchmarks and a representation formalism to aid planner comparison and evaluation
3. emphasizing new research issues and directions
4. promoting applicability of planning technology.

Commentary on the IPC websites also stresses that "The real goal of the competition is to make as much data as possible available to the community. Participants can choose to attempt relatively small subsets of the problem collections while still providing valuable data and other input into the event."

These goals are sufficiently broad to generalize as possible templates for a scheduling competition. We next discuss each of the IPC goals in turn and their suitability to a scheduling competition.

1. Analyzing and advancing the state-of-the-art

The IPC events have been formulated such that the competition domains and problems both explore limits of longstanding planner capabilities and formalize representations of emerging capabilities. To provide a measure of advances made year-to-year on a given problem type each IPC typically includes one or more domains from previous competitions that planners performed poorly on or proved especially challenging (e.g. *freecell* domain in IPC-2 & 3, *settlers* in IPC-3 & 4, *rovers* in IPC-3 & 5, and *satellite* in IPC-4 & 5).

Support for the idea that that the IPC has served to advance the state-of-the-art can be found in an overview of the problem domains, the development of PDDL, and the performance of the participating planning systems. Each successive IPC has pushed the limits for existing planning systems both in terms of the size of problems that can be handled and the extended domain modeling and nature of constraints covered (see discussion of goal 3 for examples). Early-on in the 2-year cycle for a given IPC event the organizers elicit suggestions from the planning

[1] http://ls5-web.cs.uni-dortmund.de/~edelkamp/ipc-4/

community for competition tracks and domains and may seek feedback on their own concepts for new directions. Consequently the foci of each IPC are apt to reflect both current state-of-the-art and evolutions that a cross-section of the community considers valuable. The organization supports the need for researchers to tune or extend their existing systems, or even develop new planners with the requisite capabilities by pushing to publish any needed extensions to PDDL, domain descriptions, and example problems early in the 2-year cycle.

Might a scheduling competition be expected to similarly reflect and motivate evolution of scheduling technology? An important aspect is consideration of the motivational source(s) of the advances that have been made in each field. A significant distinction between planning and scheduling to date has been the greater degree to which applications of the latter have been fielded in business, academic and governmental organizations. This may be partly attributable to the fact that for many contexts where automated planning might play a role humans have so far proven to be essential in bringing the sort of background knowledge and context to bear in selecting a sequence of actions that safely and efficiently produce a desired result. This aspect is more fundamental to planning than scheduling. As such it's not surprising that the primary impetus for most scheduling system development has come from more imminent and immediate applications, while advances in planning have been driven largely by perceived need to expand the expressiveness and breadth of the model itself in order to convincingly relieve humans of certain long-standing roles in real-world applications.

To the extent that advances in state-of-the-art are driven somewhat differently for these two technologies, it may be that the IPC formulation has provided a more valuable role for the planning community than a scheduling competition would for its subject community. Arguably many of the advances in planning technology promoted by the IPC events can be viewed as evolving the sort of sufficiently complete and robust physical world model needed to tackle the action selection problem. Without many specific extant commercial or institutional applications driving the development the role of a community-wide competition may loom larger for planning than for scheduling.

2. New benchmarks and representation formalism for planner comparison / evaluation

This has been a particularly useful aspect of the IPC for the planning community in that it's given researchers both a standard language for specifying planning domains and problems (PDDL) and a common set of problems publicly available for developers to compare their systems on. Prior to the IPC there were only isolated sets of benchmark problems and the lack of a common domain description language was an impediment to their compatibility with diverse planning systems.

The development of PDDL, the planning domain description language, has received considerable attention in the planning community. In general each successive IPC has extended this language to enable modeling more complex and expressive planning scenarios. From the PDDL limited to non-temporal, non-metric 'classical' planning problems in the first IPC, it has been progressively extended to cover such aspects as temporal planning, resource-intensive models, metric constraints, exogenous events, soft constraints on both goals and plan trajectories, and probabilistic (non-deterministic) models. Arguably even the publication of PDDL extensions for each upcoming IPC event has served to motivate planner development in directions that can handle the new language evolutions.

No such broadly recognized domain description language exists for scheduling and it's plausible that comparisons of scheduling paradigms across diverse problems would be facilitated by comparable development of a 'SDDL'. Across the diverse scheduling sub-communities one can find various sets of reference benchmark problems that have been publicly available for many years such as the OR library (Beasley, 1990) and the Project Scheduling Library (Kolisch and Sprecher, 1997). There is a considerable corpus of research that has employed these benchmarks to compare performance across algorithms. These benchmarks, however, have been focused more on certain classical domains that don't often include the breadth of constraints found in practical scheduling applications. As such, a useful role for a scheduling competition would be to promote a suitably expressive 'SDDL' that the broader community could embrace for formulating realistic, modern scheduling problems that are more closely representative of applications that drive development in this field. This then could serve as an avenue for broad dissemination of classes of practical scheduling problems over which diverse developers of scheduling technology could compare performance.

It is beyond the scope of this paper to propose a structure for a scheduling domain description language, but it seems obvious that unlike PDDL, an SDDL would support modeling of resources as first class objects. The strengths of scheduling algorithms typically revolve around efficient handling of resources so such representational differences are likely to have significant impact on performance. Given such an SDDL, an interesting possibility would be to recast a select set of benchmark planning problems from the IPCs as scheduling problems. This could lead to some perspective on relative performance of planning and scheduling approaches and the trade-offs associated with domain modeling. Relevant planning benchmark domains of interest would be resource-centric with goal and action structure that is translatable into a problem over allocation of tasks. For example, the final section summarizing the IPC events indicates that there have been several domains in the IPC events to date that are essentially scheduling domains: in IPC-2 *schedule world* and *miconic-10* and in IPC-5 *openstacks* in the deterministic track and the *elevators* and *schedule* domains in the non-deterministic track.. In addition, the *satellite* and *settlers* domains in IPC-3 and 4 with their resource management focus have a heavy scheduling orientation and could be recast as such. Once a translation process is devised it is conceivable that an automated software tool could transform a set of such

planning problems into scheduling problems appropriate for schedulers in a competition.

3. Emphasizing new research issues in planning

The planning community, through the IPC effort, has extended planning technology in part by pushing it to cover the types of models and constraints already known to be important (and therefore addressed) in real-world scheduling environments. Examples of this include:

- extension to planning systems that can be tailored to particular domains (IPC-2, 2000)
- extension to temporal models (IPC-3, 2002)
- modeling of numeric constraints (IPC-3, 2002)
- focus on specific real-world applications: 'satellite' and 'rovers' domains in IPC-3, 2002, 'airport' (based on scenarios generated by an airport ground traffic simulation tool), 'promela' (based on communications protocols), and PSR (based resolving faulty electrical networks) domains in IPC-4, 2004, 'pathways' (finding biochemical (pathway) reactions in an organism producing certain substances), TPP (based on an active OR research topic), and 'trucks' (logistics with spatial constraints, deadlines, and preferences) in IPC-5, 2006
- non-deterministic (probabilistic) domains (IPC-4, 2004 and IPC-5, 2006)
- handling of soft constraints, partial satisfiability / oversubscription (IPC-5, 2006)

As regards a scheduling competition, we have suggested in the discussion of goal 2 above that promulgation of a common, more expressive scheduling domain language would facilitate development and disseminate of more realistic benchmark problems to disseminate via the competition. Moreover, a scheduling competition may well promote progress on several emerging and challenging real-world scheduling issues that are currently difficult for different scheduling systems to compare over:

- tasks with uncertain durations and/or outcomes (Smith, et. al. 2007,)
- scheduling/rescheduling to keep pace with execution
- distributed or multi-agent scheduling
- problem-specific strategies for handling oversubscription (Kramer, et. al., 2007, Barbulescu, et. al. 2007)
- trade-offs in schedule robustness (flexibility that can absorb some degree of unexpected outcome in execution) vs. schedule quality or utility (Policella, et. al., 2004)
- trade-offs in bias towards schedule stability vs. schedule quality or utility (Zimmerman, et. al., 2006)

4. Promoting applicability of planning technology

As discussed above, scheduling enjoys a significantly higher maturity level than planning when it comes to fielded applications. Given the visibility of the two technologies, a competition that demonstrates and promotes planning applications is likely to play a more significant role for that field than a comparable competition in scheduling. To a large extent it is already the case that applications are the driving motivation behind much of the progress in scheduling research.

5. Disseminating as much performance data as possible to the community

The organizers of various IPC events have repeatedly encouraged participation by diverse groups even if a candidate planner can only handle a small subset of the domains and/or constructs being featured in a given competition. And indeed, historically most planners participating only compete or succeed in a small subset of the tracks and domains evaluated. Since the criteria for the top awards of the IPC events typically include a heavy weighting on the coverage of all domains in a subtrack - and in some cases robustness across subtracks- it is apparently not the award ranking that motivates participation by developers of such planners.

This situation benefits the larger planning community in that, while particular algorithmic approaches and/or heuristics may not be broadly effective or their implementation sufficiently mature, a comparative assessment of their strengths and weaknesses may direct attention to specific practical applications or promising directions for improvement.

This concern is especially germane to discussion of a scheduling competition: It's likely that, at least initially, few extant scheduling systems would be competitive across many domains beyond those they were specifically designed for. The designers of a scheduling competition would do well to consider options for motivating participation and structuring the subtracks to achieve enough entrants in each to provide a meaningful evaluation.

In Summary

We have compiled and surveyed the accumulated records of the five IPC events conducted to date for possible insight as to the design and utility of a scheduling competition. The cumulative IPC experience has been broadly perceived as being positive for the planning community and we have suggested both those aspects that might be expected to also serve the scheduling community as well as others that are likely to be less useful. We advocate the development of a scheduling domain description language if a scheduling competition is to be seriously pursued, noting that this could constitute a key contribution to the research community. It would however, also entail considerable upfront effort for at least the first such event.

Acknowledgments

This research was supported in part by the Department of Defense Advance Research Projects Agency (DARPA) under Contract \# FA8750-05-C-0033, and by the CMU Robotics Institute. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA.

Appendix: Synopsis of the IPC

Each of the five IPC events conducted to date is summarized here, organized according to 4 subtopics:

1. Major chosen focal areas for the IPC event
2. Evolutions in the domain language supporting the evaluation and a review of featured planning domains
3. The planning systems that entered the event
4. Summary of the evaluation and its results

Rather than provide references for each of the many planning systems mentioned, the head of each event subsection gives a general reference or web link for relevant details.

IPC-1 :1998

For details: http://planning.cis.strath.ac.uk/competition/

Foci: Non-temporal, 'classical' planning only: 1) Time represented only in sense of sequential 'steps' 2) Resources not explicitly modeled as such 3) No metric values

Domain Language & Domains: (all classical)

PDDL 1.7 is developed for IPC-1 Domains:

- *movie* Simple goals but lots of constants, causing combinatorial problems for some planners
- *gripper* 2-gripper robot carries balls between rooms
- *logistics* airplanes transport between cities, while trucks transport within cities
- *mystery* disguised logistics domain: vehicles, cargo items, and propositional fuel levels
- *mprime* disguised logistics domain with an extra action> ability to squirt 1 unit of fuel to a neighboring node if originating node has >=2 units
- *grid* robot can move 1 sq at a time on a grid. Locked squares have to be unlocked with a key of same shape.
- *assembly* assemble a complex object out of sub-assemblies obeying a given partial order.

Run as 2 "rounds" -Round 1, 2 tracks: "Strips": Basic non-conditional preconditions and effects on actions "ADL": Actions can have context-dependent effects, preconditions can be 'quantified' -Round 2: Strips only.

Competitors: 5 planners entered: Blackbox, STAN, HSP, IPP, SGP. Only the last 2 could run ADL problems.

Evaluation & Results summary: No clear-cut winner: STAN, the fastest planner on problems it could solve did not solve as many as IPP and HSP. "Big" plans were 30-40 steps long, Maximum solution sizes exceeded 100 steps.

IPC-2 :2000

For system details: http://www.cs.toronto.edu/aips2000/

Foci: Remains largely focused on 'classical' planning: non-temporal, non-metric, but extended to include new constraints such as limited metric values. Run in 2 tracks:
1) Fully automated STRIPS and ADL planners. Ranking was based on criteria including # of problems solved in each domain, CPU time required, and the "quality" of the plan produced (which ended up being the length of plans in actions and 'steps').
2) Hand tailored planning systems –Exploit domain-specific rules, axioms, and tuned algorithms.
A planned 3rd track, "Planning with resources", ultimately was not run as a separate track and relevant problems were somewhat subsumed into the first two tracks

Domain Language & Domains: Subset of the original PDDL used with some refinement/clarification for ADL:

- *logistics world* -trucks and airplanes to move packages within and between cities
- *blocksworld* -classical block stacking domain
- *schedule world* -machine a collection of parts. Goals are mostly non-interacting, but they compete for "resources" (time on machines) and on the same part different goals clobber other goals. (ADL domain)
- *freecell* -solitaire card game from Microsoft.Windows.
- *miconic-10* -based on a sophisticated Schindler Lifts Ltd. elevator controller for moving passengers. Various constraints on movement, including priority passengers, passengers that must go non-stop, passengers that must be accompanied, max number of passengers/ elevator.

Competitors: 17 planners competed: Blackbox, FF, STAN, AltAlt, MIPs, IPP, PropPlan, HSP2, GRT, TokenPlan, SHOP, TALplanner, PbR, SystemR, BDDPlan, CHIPS (the last 8 systems had hand-tailored versions)

Evaluation & Results Summary
Fully-automated track:
logistics world: only FF, STAN, Mips, HSP2, GRT, and System R scaled to the large probs.
blocksworld: only FF, System R, and HSP2 scaled to larger problems.
schedule world: Only Mips, FF, HSP2, IPP, PropPlan, and BDDPlan could handle this ADL domain. Only FF scales to harder problems (solves hardest in ~60 cpu sec).
freecell: only STAN, HSP2, FF, Mips scale.
miconic-10: On simple strips version STAN has the edge in speed and solution length. GRT does well on both criteria. On full-ADL with constraints version TALplanner (hand-tailored) solves all problems quickly while PropPlan (no domain-specific knowledge) gets shortest solutions on problems it can solve.
Hand-tailored track:
logistics world and *blocksworld*: TALplanner dominated in speed, SHOP had edge in plan length for logistics world.
TALplanner scaled to 500 block problems, in ~ 1.5 sec.
schedule world: TALplanner fastest and shortest solutions. Note that FF generates slightly longer solutions a bit more slowly, but fully automatic.
freecell: Only TALplanner solves all problems, but solutions are often long. None of planners perform much better than fully-automatic.
miconic-10: TALplanner (hand-tailored) solves all problems quickly for full-ADL with constraints version. PropPlan (no domain-specific knowledge) gets shortest solutions on problems it can solve.

IPC-3 :2002

An in-depth description of IPC-3 competitors and results referenced below are reported in Long and Fox, 2003.

Foci: 1) Extension to temporal planning at two levels of sophistication 2) Extension to numeric constraints and fluents 3) Assess the relative effort of generating and encoding control rules for planners**.** Like IPC-2 there were 2 tracks: 1) Fully automated 2) Hand tailored planners.

Domain Language & Domains:

Extended PDDL (to PDDL 2.1) to support temporal and numeric features:

- Treatment of (a finite set of) numeric-valued fluents.
- Explicit representation of time and duration.
- Plan metric specifications as part of problem instances.

Most problems were generated as 4 domain variants: 1) Basic STRIPS 2) NUMERIC, using STRIPS and metric variables only 3) SIMPLETIME, actions have duration but domain has no other metric 4) TIME, full temporal level using durative actions with durations determined by the context of their usage. The SIMPLETIME and TIME levels had no numeric resources other than time.

Two additional, more difficult variants explored combinations of the above: 5) HARDNUMERIC -Satellite domain problems with very few logical goals. Evaluation is based on amount of data recorded rather than achieving a specified logical goal. Planners challenged via the plan metric to include data acquisition actions. 6) COMPLEX problems combining temporal and numeric features.

Domains:

depots -Combines transportation style (logistics) problem with the well-known Blocks domain.

driverLog -Involves transportation, but the vehicles must be supplied with a driver before they can move.

zeno-travel Transportation problem where people must embark on planes, fly between locations and then debark, with planes consuming fuel at different rates according to speed of travel.

satellite -Inspired by scheduling of satellite observations. Satellites collect and store data using various instruments to observe a selection of targets.

rovers -Motivated by the Mars Exploration Rover (MER). Objective involves mobile rovers traversing between waypoints on the planet, conducting a variety of data-collection missions and transmitting data back to a lander. Constraints include the visibility of the lander from various locations and the ability of rovers to traverse between particular pairs of waypoints.

4 domain variants: 1) Strips: the encoding prevents parallel communication between rovers and lander: 2) Numeric*:* rover actions consume energy. Recharging can only occur at sunny locations, requiring efficient energy management. Plan quality is based on the number of recharges required, rather than plan length. 3) Simple-time: must coordinate concurrent use of rovers given the communications bottleneck with the lander and one communication channel. 4) Time: combined the demands of the Simple-time activity durations with the Numeric version's energy level management. Recharge time depends on the charge level to be replenished. Plan quality metric is makespan, but this reflects the amount of time spent recharging, so efficient energy use is important.

freecell -(replay from IPC-2) solitaire card game

settlers -Focus on management of resources, measured using metric valued variables. Products must be manufactured from raw materials and used in the manufacture or transportation of further materials. New raw materials can be generated by mining or gathering. The objective is to construct a variety of structures at various specified locations.

UM-Translog-2 -PDDL2.1 encoding of a new variant of the UM-Translog domain (hand-coding track only), a more complex transportation domain than the previous benchmarks. Domain was introduced late in the competition and very little data was collected.

Competitors: 14 planners competed: FF, LPG, MIPS, SHOP2, Sapa, SemSyn, Simplanner, Stella, TALPlanner, TLPlan, TP4, TPSYS, VHPOP. Only 3 handled domain in all 6 variants: MIPS, SHOP2, TLPlan

Evaluation & Results Summary: Plan 'quality' metric used was limited to measures of plan length: either the number of steps or the number of distinct points in the plan at which activity occurs (e.g. "Graphplan length"). In most cases the values are identical.

Planner Rankings: IPC-3 organizers emphasize problem 'coverage', success ratio, and plan quality (i.e. length) over speed. *Fully-automated track:* 1) LPG (handled 5 of 6 variants, 87% of attempted probs were solved). 2) MIPS.
Hand-coded track: 1) TLPlan (handled 6 of 6 variants, 100% of problems attempted were solved). 2) SHOP2.
'Best newcomer' planner: VHPOP (partial order planner)

IPC-4 :2004

A JAIR special track on the 4th International Planning Competition provides details of the planning systems presented below: http://www.jair.org/specialtrack.html

Foci: 1) Developed various benchmark domains that are close to applications and diverse in structure. 2) Optimal planners that provide a guarantee on solution quality were separated from the sub-optimal planners, due to the huge runtime performance gap on most of the commonly used benchmark domains. (Seven out 19 competing systems were optimal planners.) 3) Introduced uncertainty (probabilistic action effects) to the IPC; Limited to fully observable domains with discrete probability distributions.

2 major tracks:

1. Deterministic: fully deterministic and observable (also termed "classical" planning) with separate subtracks for optimal vs. non-optimal planners. Focus was on incorporation of domains that approximate applications.
2. Probabilistic: introduces a common representation language for probabilistic planners (PPDDL), and establishes some first benchmarks and results. Primary differences relative to the deterministic track: Actions may have uncertain effects, even an optimal plan may sometimes fail, value is assessed based on action cost plus goal reward, there are no durative actions, derived predicates or functions, no separate "optimal" subtrack, and planning is not separate from execution.

Domain Language & Domains:

Deterministic track: Two new constructs added to PDDL:
1) *Derived Predicates* -predicates not affected by any action available to the planner. A predicate's truth value derives from a set of rules of the form if formula(x) then predicate(x). Example: the Blocksworld "above" predicate is derived by the rule: **if** on(x,y) OR (exists z: on(x,z) AND above(z,y)) **then** above(x,y).

2) *Timed Initial Literals* --a restricted form of exogenous events: facts that will become TRUE/FALSE at time points known to the planner in advance and independent of actions it can execute. Besides the usual facts that are true at time 0 the initial state may specify literals that will become true at time points > 0. Timed initial literals are typically represented in real world scheduling problems as time windows (within which a satellite uplink is feasible, or traffic is slow, or a seminar room is occupied, etc.).

Deterministic track domains:

- *airport* Ground traffic control at airports. Test suites were generated by exporting traffic scenarios from runs of the airport simulation tool Astras. The largest test instances are realistic encodings of Munich's airport.
- *pipesworld* Flow control of oil derivatives for a pipeline network, under various constraints such as product compatibility, tankage restrictions, and (most complex domain version) goal deadlines. Novel aspect: inserting a product into a pipeline segment may produce something entirely different at the other end.
- *promela* Detection of deadlocks in communication protocols (translated into PDDL from the Promela spec. language). Deadlocks are specified via blocked transitions and processes. Communication protocols used were the dining philosophers problem, and an optical telegraph routing problem.
- *PSR* Re-supplying lines in a faulty electricity network. Electricity flow at any time point is given by a transitive closure over the network connections, subject to the states of the switches and electricity supply devices. The domain highlights "derived predicates" in real-world applications.
- *satellite* (adapted replay from IPC-3) Collection of image data with a number of satellites. Most of the IPC-3 domain versions were re-used with the same problem suites. Two of the IPC-3 domains were extended with satellite-earth transmission time windows, the most critical aspect of the real-world problem.
- *settlers* (replay from IPC-3) Test suite could not be solved efficiently by any of the IPC-3 planners. Modification for IPC-4: removed some quantified effects by replacing with lists of non-quantified effects.
- *UMTS* Setting up applications for mobile terminals. The objective is to minimize set-up time, i.e. minimize the plan makespan. When ignoring that objective (i.e. for sub-optimal planners) the problem becomes trivial. For optimal planners, the domain is a realistic challenge.

Probabilistic track: Created a version of PDDL (termed 'PPDDL') wherein effects of actions may have discrete outcome probabilities (summing to 1.0) and probabilistic initial state literals. Probabilistic track domains:

- *blocksworld* – a probabilistic variant: Blocks may slip onto table when moved. 2 versions: 1) With unit cost per action & goal reward 2) No cost/reward. Problems featured 5, 8, 11, 15, 18, 21 blocks and a goal version.
- *colored blocksworld* -variant with colored blocks and a goal of making a color-sequenced stack. Noise: Blocks may slip onto table when moved. Problems with 5, 8 & 11 blocks & goal version in 3 colors.
- *boxworld* logistics with packages and cities, drive or fly depending on edge. Noise: Get lost driving and go to wrong neighbor. Problems with 10 boxes, 5/10/15 cities.
- *exploding blocksworld* -table and blocks with noise: First put down of block may trigger explosion, irretrievably destroying object it was placed on. Costs: Goal, may become unreachable, so must plan ahead to avoid dead end. Problems: 11 blocks. Policy: Use "sacrificial" blocks to preserve stack.
- *file world* -check destination, get/put folder, put in file, placing all files in proper folders. Noise: Destination chosen randomly when checked. Costs: Getting folder is expensive, filing is cheap. Need to reason about the need to gain information. Problems: 5 folders, 30 files.
- *tire world* -drive to reach destination, replace flat, pick up spare. Noise: Tire may go flat, requiring replacement. Costs: Unit costs, high cost for "call AAA". Must construct contingent plan to do well. Problems: 30 cities. Best policy: Drive on longer route, always carry a spare!
- *towers of hanoise* -move all disks to rod3 from rod1. Single, double disk moves. Noise: Disk may slip and be lost; for doubles, slip probability depends on location of disk 5. Costs: None -goal-only version with dead ends.
- *zenotravel* -logistics for flying planes (slow or 'zoom') between cities to reach destination. (Adapted from IPC-3) Noise: Different geometric distributions for actions. Costs: -None, goal-only version. Problems: 2 cities.

"Domain-specific" (human-tuned rules allowed) and "Domain-specific, NoTuning" subtracks were included for blocksworld, colored blocksworld and boxworld domains.

Competitors:

Deterministic track: 19 competing planners

optimal planners: BFHSP, CPT, HSP*-a, Optiplan, SemSyn, SATPLAN-04, TP4-04

sub-optimal planners: CRIKEY, FAP, Fast Downward, Fast Diag. Downward, LPG-TD, Macro-FF, Marvin, Optop, P-MEP, Roadmapper, SGPlan, Tilsapa, YAHSP

Reference Planners: FF, MIPS, and LPG planners (from IPC-3) were also run on all problems they could handle.

6 of the sub-optimal and 4 of the optimal planners only dealt with the purely propositional domains.

Probabilistic track: 10 planners fielded by 7 teams.

Simon Bolivar team: mGPT, *Purdue team:* 1. Purdue-Humans (human-written policy), 2. Classy (offline policy iteration by reduction to classification, auto-acquisition of a domain-specific policy 3. FF-rePlan (deterministic replan-from-scratch), *ANU team:* NMRDPP (exploits non-Markovian rewards) , *Michigan Tech:* ProbaPOP (partial order planning, no sensing), *Dresden U of Tech:* FCPlanner (first-order value iteration in fluent calculus; domain-specific*), UMass, MSU team*: Symbolic heuristic search., *CERT*: Explicit state enumeration and DBNs.

Evaluation & Results Summary

Deterministic track: Single problem limits: Time bound> 30 min, memory bound> 1 GB. Performance data was analyzed in terms of asymptotic runtime and solution quality performance. The stated focus was the scaling behavior of planners in the specific domains. There were few distinguishing performance results in terms of plan quality. The suboptimal planners typically produced plans of similar quality, fairly close to those returned by the optimal planners, in those (generally smaller) instances they solved. For planner comparisons, each competitor identified the criterion (makespan, number of actions, or

metric value) that their planner was optimizing over for each domain version, and assessment was based on only those planners optimizing over the same criterion.
Runtime performance of some planners was more impressive than anticipated. FF, Mips, and LPG (from IPC-3) were bested by best IPC-4 planners in most cases.
Only 4 planners ran on the resource-focused 'settlers' domain, which was reintroduced from the IPC-3 event. Only SGPlan scaled to the larger problems, solving the largest in 20 sec.
Deterministic planner rankings:

- Suboptimal Propositional Track 1) Fast (Diagonally) Downward, 2) YAHSP and SGPlan
- Subopt. Metric Temporal Track 1) SGPlan, 2) LPG-TD
- Optimal Track -- 1) SATPLAN'04, 2) CPT

Probabilistic track: Planners were evaluated by simulation -the plan validator was a server with individual planners as the clients. Planners connected to the validator, received an initial state, and returned an operator/action, continuing until a terminating condition was reached, whereupon the validator evaluated the planner's performance. Results were averaged over multiple instances of this process.
Scoring: 'quality' was a combination of expected utility and running time. In goal-oriented domains, evaluation metric was the number of trials in which a goal was reached before a time limit. In reward-oriented domains, the metric was the total reward achieved before the time limit. Problem size was limited to allow computation of an 'optimal' solution. In several domains planners compared well with optimal while in others their performance degraded significantly with problem size.
Generally the competing planners each could run on only a small subset of the evaluation domains and variants, limiting the conclusions that could be drawn. Of note: The deterministic planner (FF-rePlan) that reproduced a new plan at each state was the only planner that succeeded on all problems in all domains.
Probabilistic planner rankings:

– 'Goal-based' domains: planning without using rewards, ignoring action costs. Objective is simply to maximize probability of reaching goal 1) FF-rePlan 2) mGPT
– Domain-specific: allow human-tuned rules 1) Purdue-Humans, 2) NMRDPP + control knowledge.
– Domain-specific -no tuning: 1) FF-rePlan 2) Classy
– Blind: Produced plans must be 'straight-line'; a single, contingent-safe plan at start of evaluation (equivalent to 'conformant planning' in IPC-5) : 1) ProbaPOP
– Overall Non-blocks/box: Intent of this category was to evaluate more nuanced domains than blocksworld and boxworld. 1) UMass/MSU 2) NMRDPP
– Overall: w/ goal-reward versions 1) FF-rePlan 2) mGPT

IPC-5 :2006

References for IPC-5: http://icaps06.icaps-conference.org/
Foci: 2 major tracks:
Deterministic: fully deterministic & observable (previously also called "classical" planning). 2 subtracks*: Optimal* and *Satisficing* (sub-optimal) planning
Domain categories:

- Propositional: ADL or (compiled) STRIPS domains
- Metric-Time: PDDL2.2 features, no derived effects
- SimplePrefs: propositional domains with soft goals
- QualitativePrefs: propositional domains with preferences -soft trajectory constraints
- Constraints: Metric-Time with strong trajectory constraints
- ComplexPrefs: Metric-Time with soft trajectory constraint and/or soft goals.

Non-deterministic:
Two subtracks: 1) Conformant planning: nondeterministic problems for which planners must produce a contingency-safe and linear solution. 2) Probabilistic planning*:* Focus is on real-time decision making as opposed to complete policies. Planners were evaluated using the client/server architecture developed for the probabilistic track of IPC-4. Thus, any type of planner could enter the competition as long as it is able to choose and send actions to the server.
Domain Language & Domains:
Deterministic tracks:
Extended PDDL (PDDL 3.0) to support better characterization of plan quality: Model strong and soft problem goals and constraints on plan trajectories (constraints over intermediate states reached by the plan). PDDL3.0 can express what are termed 'oversubscribed' problems in the scheduling field, in which only a subset of the goals and plan trajectory constraints can be achieved (e.g. they conflict with each other, or achieving all of them is too costly), and where reasoning over the relative importance of goals and constraints is key The plan metric accounts for soft goals and constraints ('preferences' in PDDL3.0) via penalties for failure to satisfy each of the preferences (or a bonus for satisfying them).
Only 5 of 12 competing planners handle soft constraints: SGPlan5, YochanPS, Mips-BDD, Mips-XXL, HPlan-P.
Deterministic Domains: (5 new domains plus 2 from IPC-3 & 4, resulting in 36 variants, 978 problems):

- *TPP* (Travelling Purchaser) traveling and buying goods at selected markets minimizing combined travel and purchase costs (from OR with variants, NP-hard)
- *openstacks* combinatorial optimization problem in production scheduling (from CSP benchmarks)
- *storage* -moving and storing crates of goods by hoists from containers to depots with spatial maps. Involves spatial reasoning. All 6 PDDL-3 categories represented (e.g. soft goals, deadlines, constraints requiring compatible crates stored in proximity)
- *pathways* finding a sequence of biochemical (pathways) reactions in an organism producing certain substances
- *trucks* moving packages between locations by trucks under certain spatial constraints and delivery deadlines
- *rovers* (from IPC-3)
- *pipesworld* (from IPC-4).

Non-deterministic tracks: Employed a subset of the probabilistic PDDL from IPC-4, with small extensions.
Conformant subtrack: 6 domains- 1) adder circuits 2) blocksworld 3) coins 4) comm 5) sortnet (sorting networks) 6) uts (universal transversal sequences)
There were 4, 3, 20, 25, 15 and 30 instances per domain respectively, for a total of 97 instances.

Probabilistic subtrack: 9 domains- 1) blocksworld (replay from IPC-4) noisy version of classical blocksworld 2) exploding blocksworld (replay from IPC-4) 3) tireworld (replay from IPC-4) 4) zenoworld 5) drive 6) elevators 6) pitchcatch 7) schedule 8) random

The 9 domains had 15 problems each and 30 rounds randomly generated per instance for total of 4,050 rounds.

Competitors

Deterministic tracks

18 competing planners (4 withdrew before finals):

Optimal subtrack: CPT2 (Partial-order, causal-link planning & constraint satisfaction), FDP (CSP techniques, planning graphs), IPPLAN-1SC (integer programming), Maxplan (propositional satisfiability with problem decomposition), MIPS-BDD (symbolic planning based on BDDs), SATPLAN (prop. satisfiability -new encoding).

Reference planners (IPC-4 winners): SATPLAN'04, CPT

Non-optimal 'satisficing' subtrack: Downward-sa (heuristic search), IPPLAN-G1SC (integer programming), MIPS-XXL (heuristic search, domain compilation), SGPlan5 (problem partitioning, heuristic search), HPlan-P (heuristic search, domain compilation –TLPlan extension), YochanPS (partial satisfaction planning, heuristic search)

Reference planners (IPC-4 winners): Fast Diag. Downward (Downward'04) and SGPlan (SGPlan'04)

Non-deterministic track:

Conformant: 8 planners (from 3 teams) in the finals: Conformant-FF, POND1, POND2, POND3, kp and t0 (suboptimal), sat and sat-serial (optimal)

Probabilistic: 4 planners: FPG, FOALP, Paragraph and sfDP. For comparison FF-replan, a re-planning from scratch system was also run (based on FF)

Evalutaion & Results summary:

Deterministic tracks

Optimal subtrack: SATPLAN & Maxplan dominate –they are the only ones that scale. By-domain summary:

TPP-prop: 163 actions in ~900 cpu sec.(SATPLAN)

Pathways-prop and Rovers-prop: Half of problems solved, Max problem: 135 actions in ~1000 cpu sec (Maxplan)

Openstacks-prop: 8 of 30 problems solved, Maxplan solved largest: ~1000 cpu sec.

Pipesworld: *Propositional* vers: solved ~16 of 50 problems SATPLAN max time: ~1000 sec. *MetricTime* vers: ~6 of 30 problems solved, CPT2 max time: ~1000 sec.

Storage: *Propositional* vers: solved ~18 of 30 problems, SATPLAN, Maxplan, CPT2, FDP, Mips-BDD all comparable, max time ~1000 sec. *Time* vers: solved ~10 of 30 probs, only CPT2 competed, max time: ~1000 cpu

Trucks: *Propositional* vers: ~9 of 30 problems solved, SATPLAN, Maxplan, CPT2, FDP are all comparable. Max time: ~1000 cpu sec. *Time* vers: 1 of 30 problems solved (by CPT2)

Non-optimal subtracks:

PDDL2 tracks (no soft constraints): SGPlan5 dominates. Virtually all propositional domain problems are solved with best runtimes on the most difficult problems ranging from 2 sec for storage-prop domain, to 700 sec for pipesworld-prop, to 3000 sec. for trucks-prop domain. Second place went to Downward-sa for propositional domains and Yochan-PS for 'metric-time' domains.

PDDL3 tracks (with soft constraints): 5 planners compete. SGPlan5 dominates, with Mips-XXL a distant 2nd.

-SimplePrefs & QualitativePrefs: 4 competitors; SGPlan5 dominates, solving all problems. HPlan-P is a distant 2nd, other planners do not scale. Max time: ~900 sec.

-ComplexPrefs variants: 2 competitors, SGPlan5 dominates Mips-XXL, solving almost all problems. Max time: ~800 sec

-TimeConstraints variants: 2 competitors, SGPlan5 dominates except in pipesworld where only Mips-XXL can solve a problem.

Non-deterministic tracks

Conformant: Plans were evaluated in terms of the CPU time required to output a valid plan. Conformant-FF, POND1 and t0 produce shortest length plans on average. t0 dominates on speed.

Probabilistic: Evaluation was based on a number of episodes (distinct sets of random action results) for each problem from which an estimate of the average cost to the goal of a planner's policy was computed. The 4 competing planners were then ranked on such scores. Overall ranking based on percentage of all successful rounds and total runtime. 1) FPG, 2) FOALP, 3) Paragraph 4) sfDP.

References

Barbulescu, L., Kramer, L., Smith, S., 2007. "Benchmark Problems for Oversubscribed Scheduling", International Workshop on Scheduling a Scheduling Competition, Int. Conf. on Automated Planning and Scheduling, Sept. 2007.

Beasley, J.E. 1990. "OR-Library: distributing test problems by electronic mail", Journal of the Operational Research Society 41(11) (1990) pp1069-1072.

Kolisch, R. and Sprecher, A. 1997. "PSPLIB --- a project scheduling problem library. European Journal of Operational Research, 96:205--216, 1997.

Kramer, L., Barbulescu, L., Smith, S., 2007. "Understanding Performance Tradeoffs in Algorithms for Solving Oversubscribed Scheduling Problems", Proceedings AAAI 2007, Vancouver, B.C.

Long, D. and Fox, M. 2003. "The 3rd International Planning Competition: Results and Analysis", JAIR, Volume 20, pages 1-59.

Policella, N, Smith, S., Cesta, A., and Oddi, A., 2004. "Generating Robust Schedules through Temporal Flexibility", Proc. 14th Int. Conf. on Automated Planning and Scheduling, Whistler CA, June 2004.

Smith, S., Gallagher, A., Zimmerman, T., Barbulescu, L., Rubinstein, Z., 2007. "Distributed Management of Flexible Times Schedules", 2007 Intl conf on Autonomous Agents and Multiagent Systems (AAMAS), May, 2007.

Zimmerman, T., Gallagher, A., Smith, S. 2006. "Incremental Scheduling to Maximize Quality in a Dynamic Environment", Intl. Conf. on Automated Planning and Scheduling, Cumbria, U.K., June, 2006.